Tangible Fidgeting Interfaces for Mental Wellbeing Recognition using Deep Learning applied to Physiological Sensor Data

Kieran Woodward
N0501413

A thesis submitted in partial fulfilment of the requirements of Nottingham Trent University for the degree of Doctor of Philosophy

March, 2021

Abstract

The momentary assessment of an individual's affective state is critical to the monitoring of mental wellbeing and the ability to instantly apply interventions. This thesis introduces the concept of tangible fidgeting interfaces for affective recognition from design and development through to evaluation. Tangible interfaces expand upon the affordance of familiar physical objects as the ability to touch and fidget may help to tap into individuals' psychological need to feel occupied and engaged. Embedding digital technologies within interfaces capitalises on motor and perceptual capabilities and allows for the direct manipulation of data, offering people the potential for new modes of interaction when experiencing mental wellbeing challenges.

Tangible interfaces present an ideal opportunity to digitally enable physical fidgeting interactions along with physiological sensor monitoring to unobtrusively and comfortable measure non-visable changes in affective state. This opportunity initiated the investigation of factors that would bring about the designing of more effective intelligent solutions using participatory design techniques to engage people in designing solutions relevant to themselves.

Adopting an artificial intelligence approach using physiological signals creates the possibility to quantify affect with high levels of accuracy. However, labelling is an indispensable stage of data pre-processing that is required before classification and can be extremely challenging with multi-model sensor data. New techniques are introduced for labelling at the point of collection coupled with a pilot study and a systematic performance comparison of five custom built labelling interfaces.

When classifying labelled physiological sensor data, individual differences between people limit the generalisability of models. To address this challenge, a transfer learning approach has been developed that personalises affective models using few labelled samples. This approach to personalise models and improve cross-domain performance is completed on-device, automating the traditionally manual process, saving time and labour. Furthermore, monitoring trajectories over long periods of time inherits some critical limitations in relation to the size of the training dataset. This shortcoming may hinder the development of reliable and accurate machine learning models. A second framework has been developed to overcome the limitation of small training datasets using an image-encoding transfer learning approach.

This research offers the first attempt at the development of tangible interfaces using artificial intelligence towards building a real-world continuous affect recognition system in addition to offering real-time feedback to perform as interventions. This exploration of affective interfaces has many potential applications to help improve quality of life for the wider population.

Acknowledgements

I would like to thank all the people who have helped and supported me both academically and personally throughout the process of this PhD. In particular, I would like to mention those people who have contributed most directly to my efforts to bring this dissertation to the light of day.

First, I would like to express my sincere gratitude to Prof. Eiman Kanjo my Director of Studies. Thanks to her I launched myself into this adventure of doing a doctorate. It is her invaluable guidance, constant encouragement and useful advice that have contributed to the success of this research. Thank you, for giving me the opportunity to work on such an important research project and for helping me acquire the skills necessary to become a better researcher.

I would also like to thank my second supervisor Prof. David Brown who gave me advice about how to conduct my experiments and his invaluable help of constructive comments and suggestions throughout my PhD have contributed to the success of this research. I would also like to thank my third supervisor Prof. T.M. McGinnity who has provided invaluable insights and comments. They have pushed me to consider all possible research angles and helped me become a more rigorous and thorough researcher.

My sincere thanks must also go to all of the people, past and present, that occupied the Smart Sensing lab. Thank you for all the worthwhile discussions, support and encouragement. You have been a great help, and it has been a pleasure working besides you all.

I would also like to express thanks to all of the individuals who have helped in my research via their participation in my studies. In particular, I would like to thank members of the NICER group and the teachers who have helped participate in numerous workshops and data collection studies.

Finally, I would like to show my gratitude to my friends and family, for their endless encouragement and support; I could not have completed my PhD without them.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background and Motivation

Mental health problems constitute a global challenge that affects a large number of people of all ages and socioeconomic backgrounds [102]. The World Health Organisation (WHO) [343] defines the wellbeing of an individual as being encompassed in the realisation of their abilities, coping with the normal stresses of life, productive work and contribution to their community. Stress and anxiety are one of the most prevalent mental wellbeing problems in the UK and elsewhere, yet it is still under-reported, under-diagnosed and under-treated [310]. According to the Mental Health Foundation, about a quarter of the population will experience some form of mental health problem in the course of a year [310]. Hectic modern lifestyles contribute to daily stress and a general decline in mental wellbeing, as 59% of UK adults currently experience work-related stress [236]. This makes stress the leading cause of sickness-related absences from work, with about 70 million days lost each year at an estimated cost of £2.4 billion [236]. Mental disorders are closely associated with fear of stigma, structural barriers such as financial burden, and lack of available services and resources which often prohibit the delivery of frequent clinical advice and monitoring. Advances in technologies exhibit a range of attractive properties, which facilitate the delivery of state-of-the-art monitoring.

Computers are a ubiquitous part of life although their inability to recognise mental wellbeing states results in poor human computer interactions. Affective computing [238] is computing's relationship with human emotions, if computers can recognise their users' behaviours including emotional experiences [242] it may be possible for new devices to provide intuitive interventions that improve quality of life. Expanding upon the natural affordance of touch, Tangible User Interfaces (TUIs) describe physical computer interfaces that are able to translate user actions into input events [144]. TUIs enable the interaction of digital information through users' existing skills and knowledge with physical objects, making them more intuitive to use [317], [143]. Actions on physical rather than virtual interfaces enable new forms of multi-sensory access such as fidgeting that may help regulate stress and alleviate mental wellbeing challenges [211], [141]. Little body of work has thus far looked towards tangibles as an effective interaction paradigm to automatically recognise real-world affective state and offer on-device interventions.

A vital aspect towards measuring affective state is the development of methods that can comfortably measure the non-specific responses of the body. Poor wellbeing often results in reduced Heart Rate Variability (HRV - variation in time between heartbeats) [165] and increased ElectroDermal Activity (EDA) [362] as they directly correlate with sympathetic nervous system [344] [276] [281]. This provides new opportunities to utilise non-invasive technologies for behavioural health care in order to comfortably conduct assessments in real-time. Multimodal interactions are currently used for a wide variety of purposes such as improving communication but affective computing is an area where these interactions could have a profound impact [5] [84]. By adopting a non-invasive multi physiological sensing approach and building accurate and reliable machine learning classification models it creates the opportunity for the automatic inference of affect in real-world settings.

Previously there have only been humble attempts at developing affective tangible interfaces that have not utilised advances in Artificial Intelligence (AI) such as deep learning to recognising real-world mental wellbeing. Advances in deep learning present new opportunities for the inference of wellbeing by alleviating the need for manual feature extraction and when combined with physical manipulation tools, opens the door for new forms of natural interactions and responsive interventions. By taking advantage of these advances this thesis explores the life cycle of co-designing, developing and evaluating a new form of TUI defined as tangible fidgeting interfaces. These interfaces enable real-time and momentary access to individuals' affective state and can provide therapeutic interventions on-the-go to benefit the wider population. Figure 1.1 shows the system architecture for affective TUIs from the co-design of the devices and collection of labelled data to the classification of physiological signals.

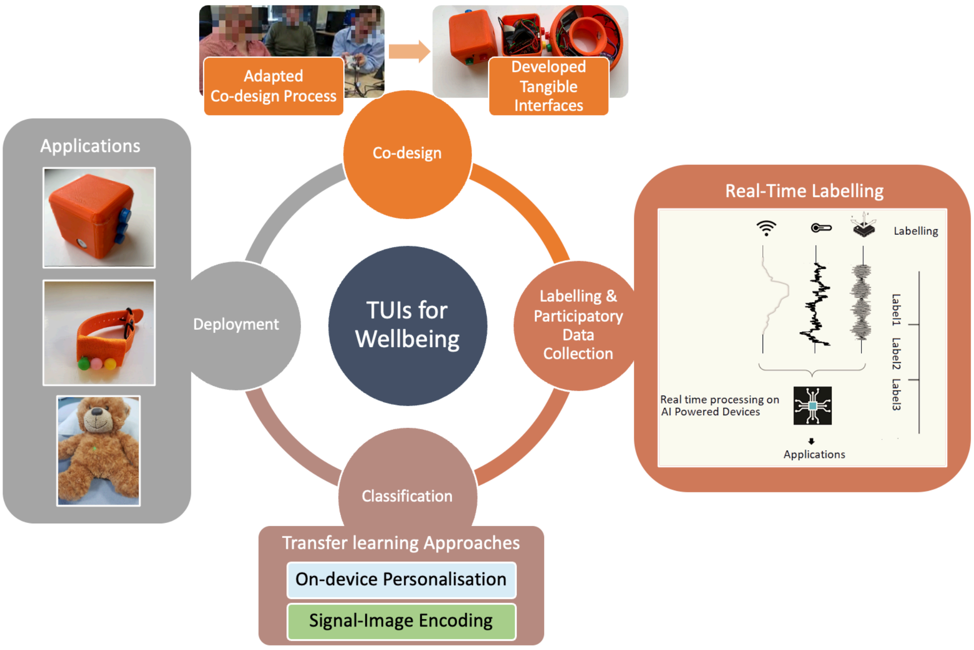


Figure 1.1: System architecture showing co-design, labelling, classification, and deployment stages.

1.2 Research Gap

The collection of physiological data in real-world environments is a challenging proposition that has resulted in the majority of previous research primarily using controlled experimental datasets, which may not transfer to the real-world domain [194], [248], [290], [368], [289], [169]. Most reported affect recognition systems for accurately measuring physiological parameters e.g. ElectroCardioGram

(ECG) usually require sticky electrodes or bands which can be burdensome. Additionally, the costs associated with these devices are usually high, limiting the potential scalability. While commercial wearables offer the potential for real-world physiological monitoring they have not been designed for the purpose of affect recognition, limiting their capabilities. Commercial wearables are hampered by poor sample rates, low accuracy [168], [30], [206], [234], [128], [218] and often forgo physical sensors such as EDA, that may help provide a better indication of health [305], [168].

Real-world labelled longitudinal data collection poses even greater challenges as it relies on multiple users continually self-reporting, while simultaneously using sensors for extended periods. Advances in edge computing are aiding on-device classification but little focus has been paid to the initial collection of labelled multimodal datasets. This results in studies that only consider subjects performing well-defined acted expressions, in a very controlled condition such as watching movie clips to elicit emotional responses [194]. Therefore, the models developed are not robust enough for real-world recognition tasks with subject variation. Most previous approaches have not considered real-world physiological data collection, making this a novel attempt at building an entire affective recognition system using multi-on-body senors and advances in AI.

Furthermore, while the use of AI to monitor affective state is not new, many previous systems have used a one-size-fits-all solution [98], [10], [340], [240]. However, due to large individual differences in physiology when experiencing different states of wellbeing personalised models may be required, especially for those with intellectual disabilities who are often overlooked. Finally, little research has considered the possibility of automatically actuating feedback on-device in real-time such as fidgeting mechanisms that may meet an unmet demand by helping people feel occupied. By combining advances in edge computing and AI it is possible to automatically activate therapeutic interventions when they are most needed.

1.3 Aim and Objectives

To help bridge the gap between human emotional experiences and computer systems this research proposes the exploration of tangible fidgeting interfaces for real-world, real-time continuous assessment. Multiple on-body sensors, such as EDA, Heart Rate (HR) and accelerometers embedded within tangible devices pave the way for continuous and non-invasive prediction of affective state.

The aim of this research is to go beyond conventional smartphone apps and wearables by developing physical interfaces that enable the real-time inference of real-world affect from non-invasive sensors including digital markers of physiology and human behaviour. Initially tangible interfaces are designed to aid the collection of real-world labelled sensor data that is required to train classification models. This thesis then explores real-world inference using a range of deep learning classifiers with an emphasis on developing personalised affective models and reducing the requirement of a large dataset for training. In order to achieve this, a set of objectives have been developed, including:

- To co-design portable tangible interfaces with the target population that combine touch and fidgeting with physiological sensors.
- To explore embedding real-world labelling techniques within the data collection component of the sensing system.
- To investigate efficient affective classification models including the exploration of personalising models, taking into account data scarcity and reducing the requirement for large datasets.
- To explore the capabilities and applications of tangible interfaces in providing real-time and continuous feedback.

1.4 Contributions

This work marks an initial attempt at developing tangible interfaces combining physical object manipulation with the monitoring of affective state through physiological sensors. The five main contributions of this work are as follows:

- The co-design and development of tangible fidgeting interfaces with end users through participatory design techniques and principles. In particular, the exploration of designs, sensors and interventions and the development of prototypes through a series of co-design workshops and focus groups. The workshops were specifically designed to engage participants who have intellectual disabilities, as mental wellbeing challenges for this target group are often misattributed to their disability [102].

- Collecting well labelled data is vital to train classification models, however labelling is an indispensable stage of data pre-processing that can be particularly challenging when applied to multimodal real-time sensor data captured from physical devices in real-world environments. Therefore, a pilot study has been conducted exploring new techniques for labelling at the point of collection running on five custom built devices, before the collection of a real-world labelled affective dataset utilising the developed techniques.

- The exploration of deep learning architectures to classify real-world mental wellbeing, some of which have not previously been used for affective modelling. A range of deep learning models including CNN, LSTM CapsNets, ResNet, Encoder and others have been trained using real-world sensor data to explore the impact different neural networks have on modelling performance.

- The development of deep classification approaches employing on-device Transfer Learning (TL) and the combination of multimodal sensor data with signal-encoded images to personalise affective models and improve performance using few labelled samples. Using the developed TL approach the process of personalising real-world affective models and improving cross-domain performance can be completed on-device, automating the traditionally manual process saving time and labour.

- The exploration of real-time, technological feedback performing as interventions to improve quality of life. Three applications are examined; sensory tools embedded within interfaces to promote fidgeting, automated on-body haptic feedback issued when poor wellbeing is inferred and wireless connectivity between interfaces to aid communication between children.

1.5 Publications

- Woodward, K., Kanjo, E., Brown, D. J., and McGinnity, T. M. (2021). Towards Personalised Mental Wellbeing Recognition On-Device using Transfer Learning "in the Wild." IEEE International Smart Cities Conference 2021.
- Woodward, K. and Kanjo, E., 2020. iFidgetCube: Tangible Fidgeting Interfaces (TFIs) to Improve Wellbeing. IEEE Sensors Journal.
- Woodward, K., Kanjo, E., Brown, D., McGinnity, T. M., Inkster, B., Macintyre, D. J., and Tsanas, A., 2020. Beyond Mobile Apps: A Survey of Technologies for Mental Well-being. IEEE Transactions on Affective Computing.
- Woodward, K., Kanjo, E. and Oikonomou, A., 2020. LabelSens: Enabling Real-time Sensor Data Labelling at the point of Collection on Edge Computing. Personal and Ubiquitous Computing.
- Woodward, K., Kanjo, E., Brown, D. J., and Inkster, B., 2020. TangToys: Smart Toys to Communicate and Improve Children's Wellbeing. Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing.
- Woodward, K., Kanjo, E., Brown, D., and McGinnity, T. M., 2020. On-Device Transfer Learning for Personalising Psychological Stress Modelling Using a Convolutional Neural Network. On-device Intelligence Workshop, MLSys, Texas.

- Woodward, K., Kanjo, E. and Brown, D., 2019. AI-powered tangible interfaces to transform children's mental well-being. 5th IEEE International Conference on Internet of People (IoP 2019).

- Woodward, K., Kanjo, E., Umair, M. and Sas, C., 2019. Harnessing digital phenotyping to deliver real-time interventional bio-feedback. Proceedings of the 2019 ACM International Conference on Pervasive and Ubiquitous Computing.

- Woodward, K., Kanjo, E., Brown, D., Kanjo, E., and Brown, D., 2019. Challenges of Designing and Developing Tangible Interfaces for Mental Well-Being. ACM CHI Conference on Human Factors in Computing Systems 2019.

- Woodward, K., Kanjo, E., Burton, S. and Oikonomou, A., 2018. EmoEcho: A Tangible Interface to Convey and Communicate Emotions. Proceedings of the 2018 ACM International Conference on Pervasive and Ubiquitous Computing and Wearable Computers.

- Woodward, K. and Kanjo, E., 2018, October. Things of the Internet (ToI) Physicalization of Notification Data. Proceedings of the 2018 ACM International Conference on Pervasive and Ubiquitous Computing and Wearable Computers.

1.6 Thesis Outline

This thesis is organised in the following chapters:

Chapter 2 provides a literature review exploring tangible user interfaces, methods to monitor affective states using physiological sensors, deep learning architectures and the application of real-time interventions.

Chapter 3 introduces the work on the design and development of affective tangible fidgeting interfaces. The co-design approach used with people who have intellectual disabilities is explored in addition to the resultant interfaces that were developed.

Chapter 4 of this thesis investigates techniques to collect and label real-world data. The development and evaluation of tangible real-time labelling techniques is explored as sensor data cannot be labelled after the point of collection. Subsequently, real-world affective data collection is completed using the developed interfaces and tangible labelling techniques.

Chapter 5 explores the classification of mental wellbeing using a range of deep learning classifiers. Furthermore, an on-device TL approach is devised to develop personalised affective models aiming to improve performance and remove the traditional challenges associated with the development of personalised models.

Chapter 6 of this thesis further investigates the use of TL with a multi-modal signal-image encoding approach to improve accuracy when using a limited dataset. This approach helps to reduce the traditionally challenging requirement of collecting a large real-world affective dataset to train deep learning classification models.

Chapter 7 discusses the potential applications of this research exploring automated interventions in addition to sensory tools and wireless communication between interfaces.

Chapter 8 concludes the work with a summary of each contribution and presents directions for future work.

Chapter 2

Literature Review

Recent advances in sensors, edge computing and machine learning have enabled the increased exploration of affective monitoring, classification and intervention. The following chapter will review the related work regarding the monitoring of human wellbeing starting with affective models, followed by methods to monitor wellbeing using apps, TUIs and sensors, then reviewing advances in deep learning to perform classification and finally exploring technological interventions. This section is adapted from [348], previously published in IEEE Affective Computing and [347], previously published in IEEE Sensors Journal.

2.1 Models of Affect

Affect, in psychology, refers to the underlying experience of feeling, emotion or mood and is an integral aspect of human life [132]. There are many aspects to monitoring affective state including measuring emotions and stress levels being felt. Where mental health conditions are clinically diagnosed [237], emotions are defined as psychological states brought on by neurophysiological changes, variously associated with thoughts, feelings, behavioural responses, and a degree of pleasure or displeasure [67]. Similarly, moods are defined as affective states typically described as having either a positive or negative valence that in contrast to emotions, are less specific, less intense and less likely to be provoked or instantiated by a particular stimulus or event [35]. In contrast, mental wellbeing is defined as a state of well-being in which an individual realises his or her own abilities to cope with the normal stresses of life and can be impacted by emotions felt [343].

Numerous psychologists have developed different theories to classify emotions ranging from small groups such as happiness and sadness [338] and pain and pleasure [212] to groups containing a larger number of emotions. There are no universal categories for emotions but the Ekman model [83] is commonly used, which comprises of 6 basic emotions: sadness, happiness, surprise, fear, anger and disgust, all of which can be distinguished through facial expressions.

Alternatively emotions can be measured dimensionally. The two most commonly used dimensions are arousal (from calm to excited) and valence (from attractive to aversive). Russell [259] describes how the arousal and valence dimensions are defined in a circle called the circumplex model of affect that can encompass all emotions, as shown in Figure 2.1.

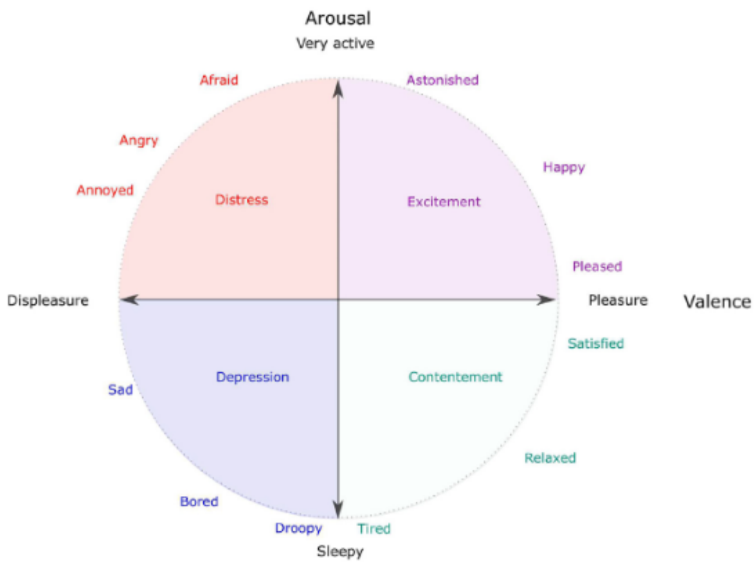


Figure 2.1: Russell's circumplex model of affect.

The Self-Assessment Manikin (SAM) Scale [38] has traditionally been used to measure valence, arousal and dominance (from submissive to dominant) using a 9-point pictorial representation of humans. This method can encourage engagement with its simpler approach, allowing for the quick assessment of affective state.

However, this approach can be challenging when collecting real-time, real-world data as it requires the immediate completion of the scale, whenever a change in emotions is experienced.

Finally, stress levels can be used to assess affective state by describing the nonspecific response of the body to any demand upon it, elicited by a stressor [277]. The stress response is mainly influenced by two aspects: first the stressor itself and second the organism's perceived ability to cope with the threat [105]. Stress can either be classified as a binary task (stress vs. no stress) [213], [241] or different levels of stress can be classified (e.g. no stress - low stress - high stress) [104].

While each of the models can be useful to capture different aspects of wellbeing, this thesis captures binary stress and categorical affective states as it provides the greatest opportunity for real-world reporting.

2.2 Methods to Monitor Affective State

2.2.1 Traditional Methods

Traditional methods used to assess mental wellbeing often use standardised clinical questionnaires, typically in the form of Patient Reported Outcome Measures (PROMs) or experience sampling [216] to understand longitudinal variability. Examples of validated questionnaires used to measure daily life stresses and symptoms include the Positive and Negative Affect Schedule (PANAS) [339], Quick Inventory of Depressive Symptomatology (QIDS) [257] and the Patient Health Questionnaire (PHQ-9) [210]. Self-reporting is used to enable people to record their emotions and stresses which can be assessed and monitored to help establish stressful triggers [298], [116]. However, self-reporting can take considerable time to assess as it must be completed over a long period to gain useful insights [297]. Also, symptom self-reporting can often be inaccurate due to poor recall; for example when a study investigated how accurately individuals self-reported the number of fruit and vegetables eaten, accuracies only ranged from 40.4% to 58% [113].

Traditionally clinical visits may also be required but these are infrequent and intermittent, representing a very small time window into patients' lives, where clinicians are challenged to decipher the possible manifestation of symptoms and trajectory. Diagnostic interviews would be performed by psychiatrists/care professionals by asking service users and their friends or family about their symptoms, experiences, thoughts, feelings and the impact they are having. Diagnostic interviews allow for a diagnosis to be made according to standard classification systems such as ICD-10 [353] and DSM-5 [11] and these are used in conjunction with a biopsychosocial formulation to construct a management plan [85], [339]. Discussions with trained experts lead to potentially identifying under-lying problems and can be used as treatment by teaching people new behaviours (e.g., to cope with stressful events). However, all of the traditional assessment methods require people to be aware of their mental health and actively seek help, which many often forego due to fear of social stigma and lack of available resources [328], [61].

2.2.2 mHealth Apps

With the high prevalence of smartphone ownership [244] access to treatment which is flexible and fits in with people's lifestyles is greatly enhanced [14]. Those at risk of mental health problems often have difficulty accessing quality mental health care [48] especially when symptoms first manifest [313], demonstrating the need for more accessible help. An Australian survey found that 76% of people would be interested in using mobile phone apps for mental health monitoring and self-management [245], illustrating the high demand for technological solutions because of their convenience and accessibility.

Many apps have been developed to modernise and advance existing practices of recording mental wellbeing. Numerous mental health diary apps are available to download, although these are effectively digital representations of existing self-reporting diaries using new techniques such as the touchscreen and monitoring notifications [360], [154], [312]. However, using a phone in public is more socially acceptable than completing a paper form, allowing monitoring to be completed discreetly in real-time, unlike paper forms which are often completed retrospectively, resulting in less accurate data being recorded [297]. A problem many apps face is the frequency for eliciting PROMs, which may under represent the true symptom's fluctuations. Given that mood is highly variable, clinically useful information is likely to be included in the daily fluctuations of mood for many cohorts suffering from mental disorders. Previous research demonstrates the possibility of eliciting daily responses to assess mental health with very good adherence over a 1 year period [314], demonstrating the feasibility of longitudinal daily PROMs. This study used engagements by two cohorts diagnosed with bipolar disorders and borderline personality disorders. More recently, chatbot apps have been developed to assess mental wellbeing, in some cases by mimicking conversation with users via a chat interface [1] thus removing the requirement to continuously self-report. A survey conducted on 5,141 participants in the age range 16-24 years showed nearly two thirds would be comfortable with a chatbot giving them a diagnosis [254]. Chatbots can utilise AI to reduce their reliance on predefined scripts and deliver individualised therapy suggestions based on linguistic analysis enhancing user engagement [66].

Figure 2.2 presents the total number of global downloads and average rating of the six most downloaded mental health apps on the Google play platform in the UK app store as of January 2019 ('7 cups', 'Headspace', 'Self-help anxiety management', 'Pacifica', 'Calm' and 'Daylio'). The total number of downloads varies widely 'Headspace', 'Calm' and 'Daylio' make up the vast majority of downloads with a combined total of 25 million, whereas the next most popular apps only amass 500,000 downloads each, showing that receiving favourable reviews does not necessarily lead to mass downloads. Apps developed by respected organisations also do not necessarily result in popularity, 'Wellmind' developed by the NHS has only been downloaded around 10,000 times and received an average rating of 3.4 out of 5 on the Google Play store, reflecting users' preference regarding usability and functionality.

Headspace currently has over ten million downloads on the Google Play store alone, underlining the immense popularity of mobile wellbeing apps. Headspace provides guided meditation and has been shown to help reduce stress by 14% [81], increase compassion by 23% [188], reduce aggression by 57% [78] and improve focus by 14% [31]. However, most of these studies were small scale with the longest period people were followed being just thirty days. Another research study reported that using the app over a six-week period resulted in no improvements

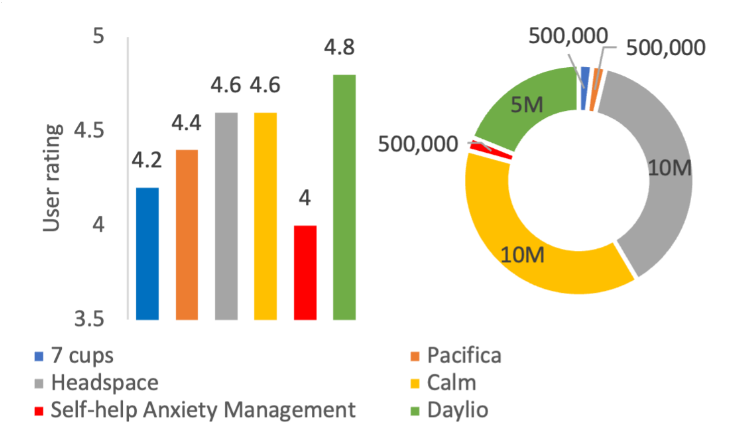


Figure 2.2: Comparison of average rating (left) and total global downloads (right) of the six most downloaded mental health apps from the Google Play store.

in critical thinking performance [222]. Additionally, there has been no follow-up after the initial studies and as some studies lasted as little as ten days it raises some concerns as to whether the positive outcomes from the app may only be apparent during an individual's initial period of use.

Additional apps have been developed by researchers that actively aim to improve mental health and wellbeing, such as mobile stress management apps that use stress inoculation training to prepare people to better handle stressful events. Studies show stress inoculation apps were consistently successful in reducing stress in participants and increasing their active coping skills [324], [323]. Grassi et al. [112] demonstrated that mHealth apps are not only capable of augmenting traditional techniques to help monitor conditions but they can also be used to educate users on more modern techniques to actively improve their mental well-being.

A smartphone app, FOCUS, has been developed to proactively ask users with schizophrenia about their mood, feelings and wellbeing multiple times each day to provide relevant coping strategies [29]. This allows the app to go beyond traditional self-reporting as it educates users on methods to help immediately after an issue has been reported, which is only possible using technology that people have continuous access to, such as smartphones. FOCUS demonstrated a reduction in positive symptoms of schizophrenia and depression, when trialled by 33 participants over 4 weeks. A common issue with mental wellbeing apps is low user engagement. However, FOCUS was used by participants on 86.5% of days, averaging 5.2 times each day over 30 days and Oiva, a mental wellbeing training app [4] was on average used every third day for 12 minutes over a 30 day period, demonstrating the possibility for mental wellbeing technologies to be highly engaging.

While apps cannot be considered as an alternative to seeking professional help some apps have been designed to work in conjunction with clinicians such as Post-Traumatic Stress Disorder (PTSD) Coach. The app allows users to learn more about PTSD, track symptoms, set up a support network and provides strategies for coping with overwhelming emotions. 10 US veterans with PTSD were assigned to use PTSD Coach independently while another 10 used the app with the support of their primary-care providers [14]. At the end of the trial, seven of the ten patients using the app with support showed a reduction in PTSD symptoms, compared with just three of the patients who used the app independently. Technological solutions used with care providers show more potential for effective treatment in the small sample trials to date although they still require users to actively seek help [201].

Pairing apps with psychiatrists' and psychologists' support has been shown to be successful resulting in a range of apps using content explicitly created by psychiatrists. Rizvi et al. [253] developed the app DBT Field Coach to provide instructions, exercises, reminders, games, videos and messages to help people cope with emotional crises. The results of that study, showed that all 22 participants used the app frequently over at least 10 days, successfully reducing intense emotions, reducing substance use cravings and improving symptoms of depression without the need to visit a clinician [253]. This app again shows the success of apps utilising psychiatrists and clinicians, although as this app only used content created by psychiatrists, it improves accessibility by removing to some extent the need to visit clinicians to access the same tools. Mobile health apps provide many advantages over traditional techniques including improved accessibility, real-time symptom monitoring, reduced cost and reduced barriers to access [80]. However, one of the main short-comings of available smartphone apps is the lack of personalised features, as many treatments and strategies have to be individually tailored [119].

2.2.3 Tangible User Interfaces

An alternative method to enhance existing monitoring techniques is through the use of TUIs. As TUIs are physical objects they rely on users' knowledge of how the physical world works for interaction [146], making them more intuitive especially for people with less digital knowledge. Matthews and Doherty [204] and Niemantsverdriet and Versteeg [220] have reported that people are more likely to create stronger emotional attachments with physical devices such as TUIs rather than digital interfaces such as apps.

Tangible devices provide a technological alternative to traditional self-reporting allowing users to report their current mental wellbeing in real-time. Emoball [95] is one such device that allows users to record their mood by squeezing an electronic ball, making users conscious of their current mood. While this device only allows users to report a limited number of emotions, participants believed mental wellbeing and education were the areas where this device could be of most use. A smaller, portable device that works similarly is Keppi [2] which allows users to squeeze to record low, medium or high pain as shown in Figure 2.3.

Figure 2.3: Keppi TUI used to record pain.

Another tangible approach to self-report is the mood TUI [273] which, as well as allowing users to record their emotions, also collected relevant data from the user's smartphone including location and physiological data such as heart rate. Participants found the use of a tangible interface very exciting, although when the device was tested with users, they felt the device was too large and they would lose motivation to continue using it for an extended period. This feedback shows the use of TUIs excites users, but the design and functionality of the interface must be prioritised. Mood sprite [25] is another handheld device developed to help people suffering from anxiety and stress by using coloured lights and an infinity mirror to assist with relaxation. The device records the time users create new sprites allowing them to be revisited much like a diary, again showing ways in which tangible interfaces can accompany traditional techniques to make treatment more accessible and user-centric. Similar to traditional self-reporting diaries, the device educates users by allowing them to recall their emotions but is more engaging with different coloured lights representing different times and moods, promoting continued use. However, a common issue with mental wellbeing tangible interfaces is that they remain largely unproven and even those that have been trialled with users such as Mood sprite have been limited to small-scale trails that lack statistical power.

Subtle Stone [21] is a similar tangible device that allows users to express their current emotion through a unique colour displayed on a stone as shown in Figure 2.4. By allowing users to set their own colours for different emotions it limits the number of people to whom users expose their emotions. Subtle Stone was tested with eight high school students in their language class with the teacher able to view the data in real-time using an app. The study showed the use of colours to represent emotions was well received with students liking the anonymity it provided along with finding it easier to use than words. Subtle Stone both allows users to communicate their emotions privately and monitor their own emotions over time, proving clear advantages over traditional self-reporting methods.

Figure 2.4: Subtle Stone TUI for communicating emotions.

A tangible interface used to detect stress in real-time without the need to self-report is Grasp, which was tested with anxious participants in a dentist's office [115]. Participants were able to squeeze Grasp whenever they felt stressed and the device detected how much pressure was exerted and displayed this data on a mobile app. Force sensors have also been used to create a tactile ball that allows for the manipulation of music by squeezing different areas of the ball along with movement detected by an accelerometer [28]. The research concluded squeeze music could successfully be used for music therapy with children as it promoted positive emotions through tactile input and music. Sensors such as force sensors have been shown to provide an intuitive method of interaction for TUIs and show the possibility for additional sensors to be utilised when educating, detecting and improving mental wellbeing, that is not possible when using smartphones or traditional techniques.

2.2.4 Physiological Sensor Measurement

This section reviews non-invasive sensors that present the most significant opportunity to assess affective state as they can easily be embedded within tangible interfaces and inconspicuously used in the real-world unlike ElectroCardioGrams (ECG) or ElectroEncephaloGrams (EEG). Features can be extracted from the sensor data to train machine learning classification models or raw data can be used within deep learning models. Alternatively, time series data can be transformed into images using Gramian Angular Summation Field (GASF), Gramian Angular Difference Field (GADF) and Markov Transition Field (MTF) [336]. These signal-image encoding techniques have previously been used to improve classification accuracy across a range of time series data including human motions and figure shapes, improving accuracy [357], reducing Mean Squared Error (MSE) by 12.18%-48.02% [336] and outperforming state of the art time series classification methods such as dynamic time warping and ResNet by 1.96%-10.13%.

2.2.4.1 Heart Rate (HR)

HR sensors are commonly used within wearable computing systems as they can be embedded within a wide range of devices due to their small footprint and provide insights into the autonomous nervous system. Similarly, HRV is commonly used within affective computing as it is the variation in time between heartbeats and further represents autonomic nervous system activity. Low HRV has been shown to be correlated with impaired parasympathetic activity, higher anxiety, a variety of depression disorders [320], [110], [217] and higher stress [250]. It is possible to measure HR and HRV using electrocardiograms [57] but in 1997 it was found that finger pulse amplitude decreased significantly during mental tasks [106] leading to HRV being accurately measured using PhotoPlethysmoGraphy (PPG). This is easier and more cost-effective to use than ECGs as it only requires one contact point.

There have been three main forms of PPG developed: transmitted, reflected and remote. Transmitted signals are most commonly used in medical monitoring, remote signals often utilise cameras to measure changes in skin colour and reflected signals measure the amount of backscattered light from an LED using photodiodes above the skin voxel where each cardiac cycle appears as a peak in the PPG signal. Reflection PPG is the the smallest and most convenient method to measure HR and HRV within tangible interfaces [198].

2.2.4.2 ElecotroDermal Activity (EDA)

EDA is often used to train affective models to classify mental wellbeing as it directly correlates to the sympathetic nervous system which controls rapid involuntary responses to dangerous or stressful situations [276]. To measure EDA the resistance between two electrodes is measured, most commonly where there is a high density of sweat glands such as the palm and finger [58]. Alternatively, near-infrared spectroscopy can be used to measure oxyhemoglobin and deoxyhemoglobin enabling the inference of stress with similar levels of accuracy as EDA [306]. However, near-infrared spectroscopy cannot be used to collect data in the real-world due to its large size and placement on the forehead.

2.2.4.3 Motion

Motion data measures movement through accelerometers, gyroscopes and magnetometers and can be used in addition to physiological sensor data to monitor mental wellbeing. Previous work has attempted to use motion data to classify emotions achieving 81.2% accuracy across 3 classes [367]. However, similar work has reported lower levels of accuracy when inferring emotions from motion data alone ranging from 50% to 72% [249] [225] [137]. Instead of using motion data alone as previous studies have researched, there is potential for it to be combined with physiological data to assist with the classification of mental wellbeing.

2.2.4.4 Labelling Mental Wellbeing

Deep Neural Networks (DNNs), are attracting more and more attention as a breakthrough in the advance of AI, showing high potential to accurately classify sensory data. However, in order to train DNNs, vast quantities of sensor data must be first collected and labelled.

The techniques used to label data vary vastly depending on the data type, as images can be labelled offline using an automated process based on clickthrough data, greatly reducing the effort required to create a labelled dataset [315]. Additionally, crowdsourcing tools have been developed that enable users around the world to label images [258], [109] and audio [179], [228], [316] while text can be automatically labelled using the hashtags and emojis contained within posts [70] or using natural language processing and machine learning [193]. Crowdsourcing labels [322] is not possible with time series data which has to be labelled online, in real-time, at the point of collection due to the nature of the data.

In pervasive sensing there are three data collection methods [360], these are: 1) Passive data sensing using smartphones or other sensors to record unlabelled data in the background [360], often used to collect weather [175], health [156] [5] and environmental data [360], 2) Active data sensing enables users to label the data in real-time through self-reporting, often used to report wellbeing or physical activity and 3) Hybrid data sensing combining both passive and active data collection as it involves users actively labelling the passive sensor data that is recorded in the background [154].

With sensor data researchers often manually label the activity participants undertake [308] which typically prevents the collection of in-situ data as it requires the researcher to continuously video participants' activities so that they can be labelled offline. Alternatively, sensor data can be labelled using a hybrid approach where the sensor data is continuously recorded and the user occasionally records a label against all or part of the previously recorded data. The labelling of human activities increasingly relies on hybrid data collection techniques to continuously record accelerometer data as well as enable users to self-report their current activity [178], [176].

Smartphone applications are becoming increasingly popular to label sensor data as they provide a familiar, always accessible interface for users [360]. However, it is not possible to use smartphones to collect data when additional sensors that are not embedded within smartphones are required e.g. EDA or HRV. It is possible for a combination of a smartphone application (for labelling) and tangible interfaces (for sensory data collection) to be used but this increases the complexity by forcing users to use two devices, along with requiring a continuous stable wireless connection between the devices.

2.3 Deep Learning Architectures

When developing a reliable wellbeing classification model it is imperative to find the best classifier. A variety of classification methods have previously been employed in the affective computing domain for classifying physiological data. Advances in deep learning have resulted in the capability to classify raw sensor data over-coming the laborious process of manual feature engineering and presenting the extracted features to a statistical learner. There are two main neural network types: Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). CNNs and RNNs are structurally different and are used fundamentally for different purposes. CNNs have convolutional layers to transform data, whilst RNNs essentially reuse activation functions from other temporal data points. This section discusses different deep learning architectures that demonstrate potential to classify physiological signals.

2.3.1 1 Dimension Convolutional Neural Network (1D CNN)

CNNs are feed forward networks that are constructed of numerous layers including an input layer, an output layer and a hidden layer that includes convolutional layers that make use of a set of learnable filters, pooling layers, fully connected layers and normalisation layers, as shown in Figure 2.5. 1D CNNs utilise a feed-forward structure like 2D CNNs, although 1D CNNs process 1 dimensional patterns with 1-dimensional convolution operations, while 2D CNNs processes 2 dimensional patterns with 2-dimensional convolutions.

In general the training input data can be represented as $x = [x_1, x_2, ...x_j]$, where the number of training samples is j and y is the output vector [121]. When σ is the sigmoid activation function, w_1 and w_2 are weight matrices between the input and hidden layer and the hidden and output layer respectively. Finally, b_1 and b_2 represent the bias vectors of the hidden and output layer respectively [364]:

$$h = \sigma(w_1 x + b_1] \tag{2.1}$$

$$y = \sigma(w_2 h + b_2) \tag{2.2}$$

Each convolution involves sliding a filter over the time series data although unlike images where CNNs are traditionally used, the filters of a 1D CNN exhibit only 1 dimension. A general form of applying the convolution for a time stamp t is given in the following equation:

$$C_t = f(\omega X_{tl/2:t+l/2} + b) | t[1, T] \tag{2.3}$$

Where C denotes the result of a convolution applied on a time series X of length T with filter ω, bias b and a non-linear function f such as the Rectified Linear Unit (ReLU). Weight sharing enables the same convolution to be used to find the result for all time stamps, allowing filters that are invariant across the time dimension to be learned.

Pooling is then performed. This can include local pooling such as max pooling where a sliding window aggregates the input data reducing it by length T. Alternatively, global pooling is where the data is aggregated over the entire dimension

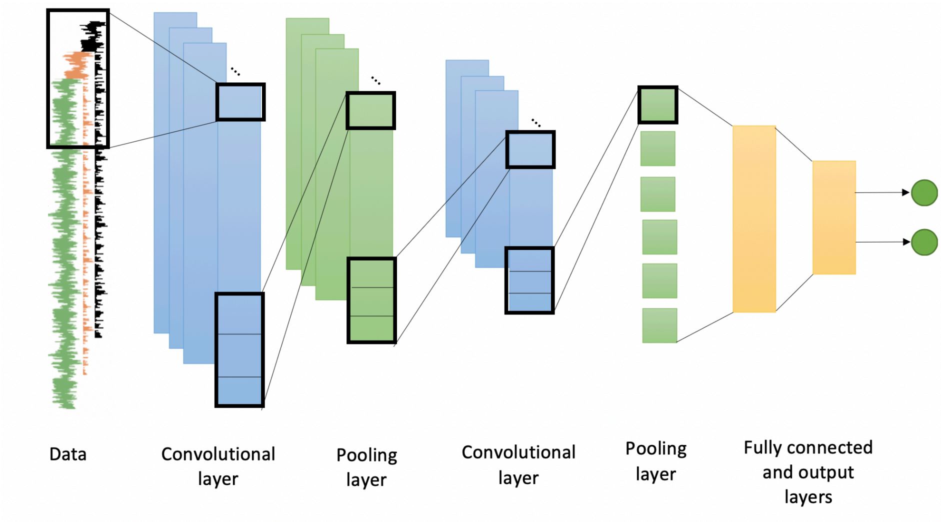


Figure 2.5: 1D CNN architecture.

resulting in a single value. In addition to pooling layers, normalisation layers can be used to help a network converge quicker.

Batch normalization can then be performed to normalise the inputs of each layer so they have a mean of 0 and standard deviation of 1. This enables the models to train quicker, allows for higher learning rates, makes the weights easier to initialise [142] and prevents the internal covariate shifting across one mini-batch training set [142]. The final layer takes the result of the convolutions and outputs a probability distribution using the SoftMax activation function.

2.3.2 Long-Short Term Memory (LSTM)

LSTM [130] networks are a specific kind of RNN where the LSTM cells serve as the memory units through gradient descent making them capable of learning long-term dependencies. LSTM cells use input (I), forget (f) and output (o) gates to regulate the flow of information, as shown in figure 2.6, helping remove the vanishing gradient problem faced by traditional RNNs.

The LSTM network uses the sigmoid forget gate layer to evaluate the inputs $ht1$ and xt and then outputs a value between 0 and 1 in the cell state to determine

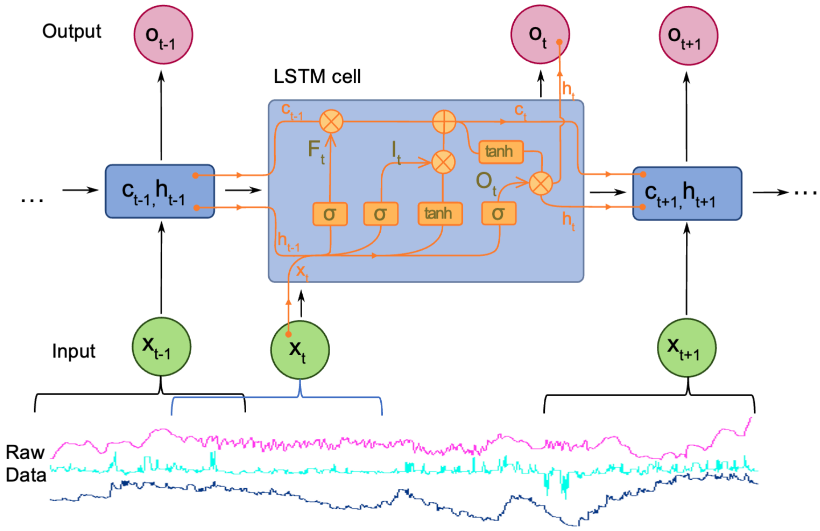


Figure 2.6: LSTM cell showing input vector (X), cell output (h), cell memory (c) and input (I), forget (f) and output (o) gates [200].

the data to forget.

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \tag{2.4}$$

Next the cell decides what new information to store in the cell state. First the input gate decides the values to update and then a tanh layer creates a vector of values, $\tilde{C}_t$.

$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] + b_i\right) \tag{2.5}$$

$$\tilde{C}_t = \tanh\left(W_C \cdot [h_{t-1}, x_t] + b_C\right) \tag{2.6}$$

The old cell state $\tilde{C}_{t-1}$ is now updated into the new cell state $\tilde{C}_t$. The old cell state is multiplied by by (f) to forget the data previously determined and then $i_t * \tilde{C}_t$ is added.

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{2.7}$$

A sigmoid layer then decides what parts of the cell state to output. Tanh is applied to the cell state and it is multiplied by the output of the sigmoid gate.

$$o_t = \sigma\left(W_o\left[h_{t-1}, x_t\right] + b_o\right) \tag{2.8}$$

$$h_t = o_t * \tanh\left(C_t\right) \tag{2.9}$$

Finally, a SoftMax layer follows the LSTM cell using a cross entropy loss function to produce an output prediction vector from the classes.

2.3.3 Gated Recurrent Unit (GRU)

The use of GRU cells is becoming an increasingly popular RNN due to its simpler design using only two gates; a reset gate and an update gate rather than the three gates used by an LSTM as shown in Figure 2.7. The use of a GRU cells can significantly reduce the time required to train models because of its simpler structure exposing the full hidden content to the next cell. GRU models have also been shown to out-perform LSTM networks when there is a smaller training dataset but LSTM models remember longer sequences than GRU models, outperforming them in tasks requiring modelling long distance temporal relations [152] [153] [358] [60].

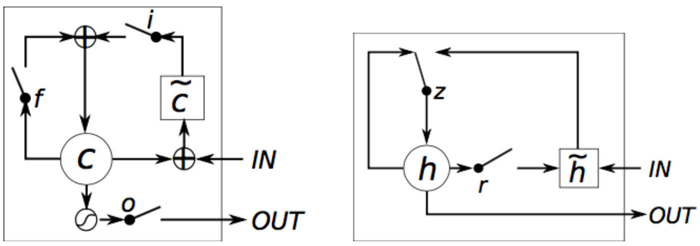


Figure 2.7: Comparison of LSTM (left) and GRU (right) cells [334].

2.3.4 Capsule Network (CapsNet)

CapsNets [129] are comprised of capsules, where each capsule encompasses a group of neurons in a layer which perform internal computations to predict the presence and instantiation parameters. Recently a 1 dimensional CapsNet has been introduced [301] that aims to preserve hierarchical spatial relationships, helping to learn faster and use fewer samples per class.

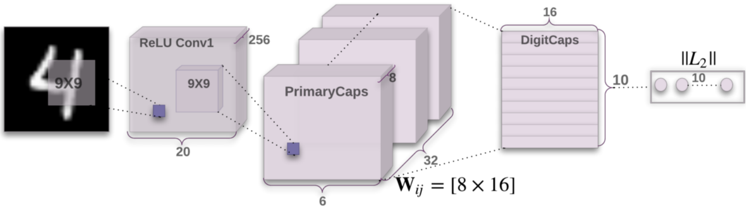


Figure 2.8: CapsNet encoder architecture [261].

The first 3 layers within the CapsNet are to encode, and the second 3 are to decode. The first layer is a traditional convolutional layer followed by a PrimaryCaps layer as shown in Figure 2.8. The PrimaryCaps layer contains primary capsules who take basic features detected by the convolutional layer and produce combinations of the features. Next, the DigitCaps layer accepts inputs from all of the capsules in the previous layer. Non-linear activations at both the Primary and DigitCaps layer are provided by the squash function. Connections between these two layers are dynamic and are governed by dynamic routing.

$$L_k = T_k max(0, m^+ - ||v_k||)^2 + (1 - T_k) max(0, ||v_k|| - m^-)^2 \tag{2.10}$$

Dynamic routing allows weights to decide which higher level capsule the current capsule will send its output to [261]. This is done by lower level capsules sending their input to higher level capsules that "agree" with the input. Two CapsNet loss functions are then used, as shown in 2.10, to equivariance between capsules and calculate the correct DigiCap. Finally, three fully connected layers decode the vector from the correct DigitCap and provide the output of the network as a vector.

2.3.5 Residual network (ResNet)

A proposed architecture from [337] is a ResNet. The network is composed of three residual blocks followed by a global average pooling layer and a final SoftMax classifier using a cross entropy loss function, whose number of neurons is equal to the number of classes in the dataset. Batch normalisation and ReLU activation function follow. Each residual block is comprised of three convolutions whose output is added to the residual block's input and then fed to the next layer. The fundamental feature of a ResNet is a linear shortcut to link the output of a residual block to its input, enabling the direct flow of the gradient through the connections and removing the vanishing gradient problem [123].

2.3.6 Time Warping Invariant Echo State Network (TWIESN)

TWIESN [307] is another RNN. For each element in an input time series, the reservoir space is used to project this element into a higher dimensional space. Then for each element, a ridge classifier [131] is trained to predict the class of each time series element. During testing, the ridge classifier outputs a probability distribution over the classes in a dataset. Then the posteriori probability for each class is averaged assigning a label for each test set where the averaged probability is highest.

2.3.7 Encoder

An Encoder network [279] is a hybrid CNN [337] where the global average pooling layer is replaced with an attention layer, enabling invariance across all layers. The first three layers are convolutional. Each convolution is followed by batch normalisation and then a Parametric Rectified Linear Unit (PReLU) [122] activation function. The output of PReLU is followed by a dropout layer and a final max pooling layer. The third convolutional layer is fed to an attention layer [19] that enables the network to learn the most important aspects for classification. Finally, a SoftMax classifier with a cross entropy loss function is used to produce a predicted label from the available classes.

2.3.8 InceptionTime

InceptionTime [91] is an ensemble of CNN models, inspired by the Inception-v4 architecture [302]. The Inception network uses a cross entropy loss function and contains two different residual blocks comprising of three Inception modules as shown in Figure 2.9. Each residual block's input is transferred via a shortcut linear connection added to the next block's input, enabling a direct flow of the gradient and removing the vanishing gradient problem.

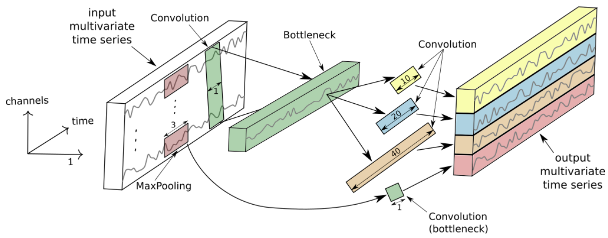


Figure 2.9: The Inception module of InceptionTime [91].

The first component within the Inception module is the "bottleneck" layer. This layer performs an operation of sliding filters of length 1 with a stride equal to 1, allowing for longer filters than ResNet. The next layer involves sliding multiple filters of different lengths simultaneously on the same input data. The output of the sliding MaxPooling window is then calculated and the output of each independent parallel MaxPooling layer is concatenated. By training the weights of multiple inception models using filters of varying lengths, the network is able to extract latent hierarchical features.

2.3.9 Multi Channel Deep Convolutional Neural Network (MCDCNN)

MCDCNN is a CNN where the convolutions are applied independently on each dimension [369]. Each dimension of input data passes through two convolutional layers with ReLU as the activation function followed by a MaxPooling operation. The output of the second convolutional layer for all dimensions is concatenated over the channels axis and then fed to a fully connected layer with ReLU as the activation function before the SoftMax classifier.

2.4 Mental Wellbeing Classification

Deep learning algorithms have generated great impact in affective modelling. RNNs relying on LSTM are especially valuable for use with sensor data as they are fundamental in distinguishing similar data which differ only by the ordering of the samples which can often dictate differences in mental health [229]. CNNs have traditionally been used to classify images and speech, however their application has been expanded to classify raw sensor data [202], [162]. This section explores the different methods used to classify mental wellbeing including mobile apps, physiological sensors including the use of TL and the exploration of real-world classification.

2.4.1 Mobile Apps

Apps have been shown to enhance traditional PROMS-based assessment techniques and by utilising the sensors within phones the capability of apps can be further enhanced as they may provide a more holistic picture using passively collected data. Smartphones are capable of collecting a vast amount of data such as location, motion and phone use which can result in many features being extracted to train machine learning algorithms. It is possible to use the data collected from smartphones to determine emotions with a 70% accuracy utilising machine learning to process the data [365]. Automatically inferring emotions based on smartphone use is extremely valuable in determining mental wellbeing and can provide new clinical insights from passively monitoring users' behaviour. Furthermore, an app has been used to monitor social media use (Facebook and Twitter) and capture ground truth mood data [182]. The results demonstrate the ability to infer real-world mood from social media data with up to 95.2% accuracy showing the possibility of monitoring wellbeing through online activity.

In addition to using the phones' sensors to detect mental wellbeing, it may be possible to use the phone's touchscreen to sense stress. Using an infrared touchscreen to measure HR it was possible to recognise stress with accuracies of 87% and 96% across two tests, a vast improvement upon previous touchscreen based stress detection [366]. However, infrared touchscreens are rarely used especially within smartphones whereas the capability of measuring stress through capacitive touchscreens could have significant impact.

Smartphone apps have also been paired with wrist-worn sensors to infer mental wellbeing by allowing a high magnitude of data to be collected [270]. The collected data was expressed using 15 multimodal features ranging from physiological data such as skin conductance to phone usage data such as screen time duration. The 15 sets of features were then trained with a variety of classifiers and the accuracy of the different features were examined for each classifier. The system was capable of detecting binary stress with over 75% accuracy, with some of the features such as increased acceleration during sleep and high evening phone use being more beneficial than others in determining stress.

Similarly, a wrist sensor along with a mobile app and a self-reported PHQ-8 and PHQ-4 depression scores were used to quantify depression symptoms in 83 undergraduate college students across two 9-week periods by measuring phone use, heart rate, sleep and location [335]. The study concluded students who reported they were depressed were more likely to use their phone at study locations, have irregular sleep, spend more time being stationary and visit fewer places. They demonstrated that they could automatically detect depression with 69.1% precision when evaluated against the PHQ-4 depression sub-scale [171], this could potentially be improved if additional physiological sensors were included. In addition to physiological sensors, location could be used to assess mental wellbeing as movement patterns and uncertainty in visits has been shown to be predictive of the Quick Inventory of Depressive Symptomatology (QIDS) [232].

Alternatively apps can be developed for smartwatches. BreathWell, [329] which has been developed for Android Wear smartwatches has been designed to assist users in practising deep breathing to reduce stress from PTSD, although the app has limited functionality to determine stress as it only uses the user's heart rate. Despite the limited functionality, all seven participants believed the app could help them, although the extent of the trial was extremely limited. These studies demonstrate the potentially powerful combination machine learning, sensors and mobile apps provide to automatically determine stress levels.

2.4.2 Multimodal Physiological Sensors

There are numerous sensors that when combined with sufficiently trained machine learning classifiers can go beyond mobile apps to assess mental wellbeing in real-time. Non-invasive physiological sensors such as EDA and HRV present the most significant opportunity to assess mental wellbeing [281] with EDA alone detecting stressed and cognitive load with 82.2% accuracy [276]. A CNN has been trained to classify four emotions; relaxation, anxiety, excitement and fun using EDA and blood volume pulse signals [202]. The deep learning model outperformed standard feature extraction across all emotions achieving accuracies between 70-75% when the features were fused. Furthermore, when EDA sensors were paired with accelerometers to classify human activity and stress using logistic regression, 91% accuracy was achieved [372]. EDA, blood oxygen and HR have also been used to classify five emotions using random forests achieving 74% accuracy [340]. When using 30 statistical features from EDA data to infer six emotions; surprise, fear, disgust, grief, happy and angry 78.72%, 73.37%, 70.48%, 62.65%, 62.52% and 44.93% accuracy was achieved for each of the six emotions respectively [354]. Finally, when EDA and ECG signals were combined to infer happy, sad and neutral emotions accuracies ranged from 90.12% for happy-neutral to 93.32% for happy-sad [68].

Both EDA and HRV were used in a wearable device to measure stress during driving [124]. The wearable device took measurements over a 5-minute period to detect stress levels with an accuracy of 97.4% and found that HRV and skin conductance are highly relatable making them extremely useful in detecting mental state. A minimum redundancy maximum relevance selection algorithm has demonstrated the capability to select the optimal HRV features to train machine learning classifier, achieving 84.4%, an increase of 10.8% over the sample base SVM [101]. The ability to use sensors to measure HRV and skin conductance allows for small devices to accurately determine stress levels in real-time and should be further utilised to detect stress, anxiety and mental wellbeing. However, physiological signals do not account for the context in which the devices are used as the context can play a significant role in the users' perceived stress levels meaning additional environmental sensors may also be required [300].

EDA and HR signals have also been used to infer stressed and relaxed stated using K Nearest Neighbour (KNN) and Fisher discriminant analysis achieving 95% accuracy stating that 5-10 second intervals are suitable for real-time stress detection [73]. A controlled stressor experiment with 166 participants completing puzzles to collect EDA and HRV signals when experiencing high, medium and low stress has been conducted. A finite state machine classifier achieved 0.984, 0.970 and 0.943 F-measure scores for high, medium and low levels of stress respectively [203].

Another non-invasive sensor that has previously been used to detect stress is skin temperature as it can indicate acute stressor intensity [127]. One study [164] used a wearable device that contained multiple sensors including skin conductance, skin temperature and motion and provided it to 6 people with dementia and 30 staff in a nursing home for 2 months. The device aimed to automatically detect stress and categorise it into one of five levels, the accuracy for each of these levels varied from 9.9% to 89.4% showing an extremely wide variation. This was due to the threshold setting: when it was raised, fewer events were classified as stress because of the harder criteria, in turn, increasing precision. Furthermore, when inferring six emotions HR, EDA, skin temperature and environmental temperature were used within a wearable glove and chest belt [192]. Using KNN, direction finding and multiple back-propagation classifiers with four features, emotions could be classified with accuracies between 72.3% and 84.1%.

Stress can also be detected from brain activity using EEG [59] as Khosrowabadi et al. demonstrates using eight channels to classify students' stress during exams with over 90% accuracy [162]. A CNN with channel selection strategy, where the channels with the strongest correlations are used to generate the training set, has also been used to infer emotions from EEG signals [247]. The model achieved 87.27% accuracy, nearly 20% greater than a comparative model without channel selection strategy. Similarly raw EEG signals have been used to train a LSTM network achieving 85.45% in valence [8] and to classify stress in construction workers with 80.32% accuracy using a gaussian support vector machine model [149]. Dynamical Graph CNNs have also been used to infer emotions from EEG signals achieving 90.4% accuracy using subject dependent models and an average of 79.95% accuracy with 15 subject independent models [290]. However, the inconvenient placement and cumbersome nature of EEG sensors result in limited use in real-world environments.

Another device aimed to detect stress using ECG, EDA and ElectroMyoGraphy (EMG) of the trapezius muscles [344]. Principal component analysis reduced 9 features from the sensor data to 7 principal components. 18 participants completed three different stressors; a calculation task, a puzzle and a memory task with a perceived stress scale questionnaire completed before and after each task. The principal components and different classifiers were used to detect stressed and non-stressed states with an average of almost 80% accuracy across the three tests compared with the questionnaire results. However, this study only detected two states; stressed and non-stressed and was conducted in a controlled environment so it is not known how accurate it is in real-world setting as physiological signals can be affected by factors other than mental wellbeing.

Furthermore, LSTM networks have been used to classify sensor data including EDA, skin temperature, accelerometer and phone usage data to infer stress. The LSTM model achieved 81.4% accuracy outperforming support vector machine and logistic regression models [319]. LSTM networks have also been used to classify emotions from EEG signals with 81.1% accuracy when using the context correlations of the feature sequences [356]. A CNN and LSTM have been combined to allow raw data to be classified more accurately [155], [156]. This deep learning approach is capable of using raw data to automate the feature extraction and selection. This approach to classifying emotions from physiological, environmental and location data outperformed traditional multilayer perceptrons by over 20%. The ad-hoc feature extraction by the CNN matched or outperformed models with the features already extracted showing the clear advantages of using deep learning approaches.

Sensing devices are increasing in popularity with advancements in physiological and environmental sensors resulting in cheaper and smaller devices promoting extensive use. The ability to pair machine learning algorithms with sensors presents an enormous opportunity allowing for mental wellbeing to be detected with accuracies exceeding 90% [162], [124]. While AI has enormous potential in classifying affective states, it does present its own set of challenges as a large amount of labelled data is required to train the models accurately.

2.4.3 Transfer Learning (TL)

TL [233] is a common approach in machine learning to mitigate the problem of scarcity of data. Caruana [54] introduced multi-task learning that uses domain information contained in the training signals of related tasks. It is based on the ability to learn new tasks quickly even without many samples, by relying on previous, similar samples. TL involves pre-training a model on a task which is similar to the target task but has significantly more training data available and transferring the learned knowledge.

CNNs are commonly used in TL as they are initially trained on a large dataset and then the last fully-connected layer is removed and the model is further trained on a smaller target dataset. A pre-trained CNN alleviates the need for a large dataset while simultaneously decreasing the time required to train the model. The premise of TL is to improve the learning of a target task in three ways [311]: (1) improve initial performance, (2) sharp performance growth, (3) potential higher training performance.

TL has most commonly been used for object recognition [227], human activity recognition [271] and speech recognition [333] although it could be used to address the challenges of affect recognition. By using a pre-trained model from a different domain and transferring the learned knowledge to the new domain it is possible to improve modelling performance while training using few samples.

Previously, large ImageNets have been used to developed pretrained networks such as VGGNet [285], Inceptionv3 [303] and mobileNetv3 [139] that contain pre-trained object classification models. The pre-trained CNN models were employed to compute mid-level image representations for object classification in PASCAL VOC images [88], leading to significantly improved results. TL has enabled the possibility to easily train new models in the visual domain using pretrained networks, however, sensor data is not always easily visually interpreted [280].

TL using physiological signals has previously been used to detect driver status [177] and seizures [72]. Furthermore, an inter-subject TL approach has been used with ECG signals to infer mental state achieving 79.26% compared with a baseline of 67.90%, demonstrating the potential for TL to improve affective model performance with small physiological datasets [89].

The possibility for TL to be used to personalise affective models has previously been explored and has helped personalise EEG signals, improving model accuracy by 19% [368] and 12.72% [185] while also reducing the amount of data required to train the models. TL can be used to help alleviate scarce data as by using decision trees, data from similar subjects can be used to improve accuracy by around 10% although if data from dissimilar subjects is used it can have a negative impact on the model accuracy [205]. To ensure negative TL that degrades the performance of the model does not occur, a conditional TL framework has been developed that assesses an individual's transferability against individual's data within the dataset. The conditional TL model identified 16 individuals who could benefit from 18 individuals data within the EEG dataset, improving classification accuracy by around 15% [189].

The inference of emotions from images and videos has also benefited from TL approaches. When using models pre-trained on the ImageNet dataset and testing using images of faces expressing seven emotions an accuracy of 55.6% was achieved compared with a baseline performance of 39.13% [219]. Additionally, audio and video have been explored to infer six emotions where the TL approach improved base line accuracy by 16.73% [230].

Another TL approach has helped increase the classification accuracy of PTSD using speech by 13.5% [22]. Similarly, a sparse autoencoder-based feature TL approach has been developed to infer emotions from speech using the FAU AiboEmotion Corpus dataset [27]. The autoencoder approach to find a common structure in a small target base dataset and apply the structure to source data improved unrated average recall from 51.6% to 59.9% with only 50 data instances used [77]. Whispered speech has also been explored to infer emotions applying three TL approaches; denoising autoencoders, shared-hidden-layer autoencoders, and extreme learning machines autoencoders. Extreme learning machines autoencoders provide good generalisation extremely fast [140], enhancing the prediction accuracy on a range of emotion tasks achieving up to 74.6% arousal [76]. Speech has also been explored to improve PTSD diagnosis using TL and deep belief networks. The TL approach improved model accuracy from 61.53 to 74.99% [23]. Furthermore, deep belief TL networks have been used to improve the accuracy of emotion recognition through speech cross-language [180]. TL for emotion recognition has also been used to infer wellbeing from text [283]. A RNN with full weight transfer where the base model was trained using a Twitter dataset to classify tweets as positive or negative valence achieved an overall accuracy of 78% for all four classes where the standard RNN achieved 72%.

2.4.4 Real-world Classification

The vast majority of previous literature has classified mental wellbeing in controlled experimental conditions. Real-world environments provide many additional challenges as there can be numerous factors that influence physiology rather than purely wellbeing such as music [166], sleep deprivation [243] and nutrition [359]. However, it is necessary to test models in real-world environments to ensure their effectiveness in every day scenarios. HRV and EDA physiological sensors present the largest opportunity for real-world mass adoption monitoring due to their non-invasive, inconspicuous nature.

Classifying stress has been explored in the real-world [103]. Using 63 features from blood volume pulse, HR, EDA, skin temperature and respiratory rate sensors, binary stress was inferred with 83% accuracy and 3 stress levels with 72% accuracy. When inferring stress in real-world environments binary classification reduced to 76% but by using accelerometer data to provide context the accuracy was increased to 92% over 1 hour periods. HRV has also been used to infer stress over periods of 3 minutes from 45 students experiencing a stress inducing oral exam. Using a decision tree classifier accuracies of around 80% were achieved [55].

HRV data from chestbelts along with audio, physical activity and communication data from a smartphone has similarly been used to infer real-world stress [214]. Using the PANAS questionnaire as ground truth labels, 13 features obtained from the smartphone data and 10 HRV features which were more distinctive than the smartphone features, 61% accuracy was achieved for high, medium and low stress classification. Similarly when using a chest band respiratory sensor and accelerometer to infer stress using an SVM, classifier accuracy declined from 89% in controlled conditions to 72% in the real-world demonstrating the challenges of inferring real-world mental wellbeing [138].

Multi-task learning has been used on real-world data with students reporting their health, stress and happiness each morning and evening [147]. The best performing models were trained using 343 features extracted from EDA, skin temperature and accelerometer sensors, smartphone logs and weather achieving 82.52%, 86.07% and 87.5% accuracy for health, stress and happiness respectively. Furthermore, 19 participants collected EDA and HR sensor data over 5 days labelling their valence and arousal whenever they experienced a change in emotions [125]. The model successfully classified 85% of the high and low energy emotions and 70% of the positive and negative emotions.

Similarly, to monitor mood at work a smartphone app and wearable embedding an ECG, EDA, accelerometer and skin temperature sensor has been developed [363]. 4 users used the system over a period of 11 days labelling their emotions every 8 hours to collect the sensor data to train the KNN, decision tree and Bagged Ensembles of Decision Trees (BE-DT) networks. When inferring the strength of 8 emotions the BE-DT model achieved the highest performance of 62.1% for generalised models and 70.6% for personalised models demonstrating the capability to infer real-world emotions albeit using a small test sample. However, there remains many challenges when developing a classification model to infer real-world wellbeing including collecting data from a large sample and collecting ground truth labels [52].

2.5 Technological Based Interventions

An area of application still in its infancy is technologies that go beyond sensing to additionally provide feedback helping to improve mental wellbeing. A variety of feedback mechanisms can be used to improve wellbeing and by combining these with real-time classification models it enables the possibility for real-time interventions to be automatically applied.

2.5.1 Biofeedback Therapy

One method to improve mental wellbeing is bio-feedback therapy; this involves monitoring a normal automatic bodily function and then training people to acquire voluntary control of that function. Nolan et al. [221] measured HRV in patients with coronary heart disease as cardiac death is more likely in these patients when stressed. The study recruited 46 patients, of whom 23 undertook HRV biofeedback involving training patients in paced breathing in order to improve their HRV and stress management. The study resulted in patients showing reduced symptoms of psychological stress and depression, proving the positive effect of biofeedback training and controlled breathing. However, further work is required to investigate whether these findings could be generalised within real-world environments.

Another study [170] used biofeedback for general stress management; this biofeedback used a game to encourage users to improve their HR and cerebral blood flow control. This study used stress focused questionnaires, a stress marker and a voxel-based morphometric analysis to determine stress, allowing the study to conclude that the biofeedback helped reduce daily stress due to the increase in regional grey matter [100]. HRV biofeedback has also been used during the postpartum period after the birth of a child. The study [174] showed that biofeedback helped improve HRV and improve sleep over the 1 month period it was used by 25 mothers. However, the lack of a control group means the study does not definitively show the improvements were due to the biofeedback training.

Biofeedback has been shown to have a significant impact in reducing stress during trials although its effectiveness in real-world stressful situations has not been proven [341]. The possibility of pairing bio-feedback training with virtual reality could allow users to practice the techniques learned through biofeedback to reduce stress in a setting they find stressful, which would demonstrate the effectiveness of biofeedback. Furthermore, biofeedback requires people to have an understanding, willingness and time to train their body to acquire voluntary control which many people do not possess. Tangible interfaces may solve many of these problems by using sensors to analyse affective state similar to biofeedback, and additionally provide automated feedback to improve wellbeing in real-time.

2.5.2 Real-time Feedback

Devices that sense and provide feedback ranging from tangible interfaces to robotics have the possibility to positively impact the wider population who may temporarily experience mental wellbeing challenges but do not seek professional help. Tangible devices offer the capability to improve mental wellbeing in real-time by pairing sensors with automated feedback.

A variety of tangible mental wellbeing devices have been produced by Vaucelle, Bonanni, and Ishii [321]. These include *touch me*; which contains multiple vibrotactile motors to provide the sensation of touch, *squeeze me*; a vest to simulate therapeutic holding, *hurt me*; a wearable device that applies a moderated painful stimuli to ground people's senses and *cool me down*; a device that heats up to ground people's senses. From the devices developed clinicians believed *hurt me* had the most potential as it could allow for the patient and therapist to better relate to one another, by having the therapist working with the class of pain the patient is experiencing psychologically and externalising viscerally. All of these interfaces have specific purposes such as *hurt me* which may be beneficial for people considering self-harming but not for people suffering from other mental health challenges. A more general device is required for people who may experience temporary mental wellbeing challenges.

It is possible to help improve general mental wellbeing using small devices with real-time interventions; one such device is Squeeze, Rock and Roll [47]. This device allows users to simulate rolling behaviours as many people do with a pen when stressed but the device gradually guides the user to reduce their movements and their stress through dynamic tactile feedback. However, while people acknowledged the device helped them relax no stress reduction was found, possibly because the device offered very little feedback. Guiding users' behaviours is a novel approach to improve mental wellbeing although possibly less effective as some people may find the action of rolling or twisting objects relaxing and is often used as a coping strategy for people suffering from mental health conditions [53].

Haptic feedback is a method of providing feedback that recreates the sense of touch through the use of motors and vibrations; this allows people to experience real sensations which can significantly affect emotions and has been shown to successfully improve mental wellbeing [332], [62]. Good vibes [159] used a haptic sleeve to provide varying feedback dependent on heart rate readings. A stress test was conducted while the sleeve used dynamic vibrations to help reduce the heart rates of the participants by 4.34% and 8.31% in the two tests compared to the control group. Doppel [17] also used haptic feedback in a wearable device that aimed to reduce stress before public speaking, measuring users' heart rates and skin conductance to determine stress. The speed of the vibration was dependant on the user's HR, providing personalised real-time feedback. When users were told they were to present a speech the skin conductance data showed users wearing the Doppel remained less stressed than the control group. This research shows that haptic feedback can have a substantial positive impact in improving mental wellbeing and is more successful than guiding user interactions. The advantage of personalised haptic feedback is clear, but more research needs to be conducted to establish the best rate and type of feedback for individual users.

A headband has also been developed that uses EEG combined with machine learning to assess stress by analysing alpha and beta waves as alpha waves decrease when stressed [278] and then uses massage therapy to reduce stress [50]. The massage motors were tested on 4 participants with 3 of these responding well to the feedback and becoming less stressed showing the possibility for massage therapy to be further utilised in stress reduction devices. However, as the device was only used by 4 participants with a 75% success rate, it is far too small to have any statistical significance with much more research needing to be conducted to prove it can be as effective as haptic feedback.

Communicating with others has a positive mental impact leading to research that remotely connects people through biofeedback. When communicating emotions, a range of output modalities have been explored including including visual, haptic and audio feedback in addition to more novel feedback modalities such as temperature, shape changing, taste and smell [184]. Haptic feedback has commonly been used due to its unobtrusive nature and ability to represent the feeling of touch [215]. Haptic feedback can further be utilised as a breathing pacer. Miri [208] found that the development of a personalised routine where users find the breathing frequency comfortable was more important in relaxing users than the placement of the feedback itself, demonstrating the benefits of haptic feedback to both passively calm users and actively aid relaxing breathing practices. Similarly, shared breathing experiences through Breeze using tactile, visual and audio feedback helped to increase the feeling of belonging between connected participants [94]. Communication with others is vital to ensure positive mental wellbeing and while feedback devices that remotely connect individuals appear to improve mental wellbeing they have only been tested in limited trials.

Many existing approaches have not classified wellbeing to issue feedback but instead mirrored physiological reactions back to users for their reflection [264]. Visual feedback in the form of flexible wrist-worn displays have been developed enabling users to view real-time representations of their physiology [318]. However, the use of a thermochromic displays resulted in the visualisations being ambiguous and potential privacy concerns as other people could easily view the wearer's emotions.

Somaesthetics has also been used as a method of combining bodily experiences with aesthetic appreciation, presenting many opportunities for expressing mental wellbeing [346]. The Breathing Light [135] created an enclosed space where light dynamically changed with the user's breathing patterns enabling deep reflection. The Breathing Light provided a more calming alternative for visual feedback but may be challenging to adopt in real-world environments. Alternatively, BrightHearts [163] is a biofeedback mobile app that changes patterns, colour and sound with the user's HR, helping people become self-aware of their wellbeing and assisting them to relax as shown in Figure 2.10. Another app aimed to improve emotional wellbeing without the use of somaesthetics, instead using momentary photography [183]. The results demonstrated photographic features such as the number of photos taken and the number of photos revisited were positively correlated with an improvement in the participant's mood showing photography as a simple real-time method to improve wellbeing.

BioFidget [186] shown in Figure 2.11 is a self-contained device that monitors HRV and allows users to train their breathing by blowing on the fidget spinner to reduce stress. Twenty participants stated BioFidget helped them feel relaxed and overall it helped the majority of users improve their HRV showing they were less stressed. Inner Flower [117] has also been developed as an ambient device that uses the user's HR and HRV to create a breathing guide using visual feedback. The results showed that breathing exercises reduced stress levels although the ambient visual feedback had no significant impact on stress.

Additionally, somatics has been used to translate EDA data into felt experiences through changes in temperature [7]. Wearer's of the device found the

Figure 2.10: BrightHearts app showing light pattern.

cooling sensation triggered by changing arousal levels to be pleasant. The use of varying temperatures as feedback is a novel method that when paired with additional sensors to better monitor wellbeing could provide subtle calming feedback that may help improve wellbeing.

A different approach to provide real-time feedback is to alert the user regarding their current mental state, allowing them to take appropriate measures such as reducing workload or taking time to relax. MoodWings [197] aimed to reduce stress through wing actuations informing users of their current stress levels. Participants wore the device on their arm while ECG and EDA readings were taken to determine stress. A simulated driving experience was undertaken by participants and once stress was detected the wing movement was manually activated. The results show that MoodWings improved the participants' awareness of their stress, but their awareness further increased their stress as shown by EDA data, resulting in the device having a negative effect on users' mental wellbeing. Overall this study demonstrated that sharing wellbeing data with users needs to be carefully considered [197].

A novel approach to provide feedback is through the use of robotics such as therapy animals which are most commonly used to reduce loneliness. One

Figure 2.11: BioFidget TUI used to promote deep breathing.

example of a robot used for therapy is Paro; a robotic seal that was designed as an easy to use robotic animal that encourages user interaction with its large eyes and soft fur [282]. Tactile sensors allow Paro to understand the location and force of users' touch allowing for the response's magnitude to be relevant to the input. Studies show Paro provided extremely effective therapy as it helped reduce stress in a day service centre for elderly adults [16], increased user's social interactions and improved their reactions to stress in a care home [282]. Paro has been shown to have a great impact in helping reduce stress in elderly adults even with its limited sensors and responses and has the potential to have a wider positive impact on people's mental wellbeing.

Figure 2.12: Paro therapy seal.

Although most therapeutic robots such as Paro target the elderly, a robotic teddy aimed at reducing stress in young children at the hospital has been developed [150]. Rather than relying upon tactile interaction like Paro, this teddy uses vocal interactions, which children preferred. The children who used the robotic teddy spent more time playing with it than the comparative virtual or traditional plush teddy, they also had more meaningful interactions and their behaviours conveyed they were emotionally attached to the bear and not stressed. Robotic interactions can have a positive impact on emotional experiences and help reduce stress in both the young and the elderly. Robotic animals could be easily adapted to incorporate additional sensors to automatically detect mental wellbeing allowing for more personalised responses to be produced.

Feedback devices aim to advance upon sensing devices by actively improving mental wellbeing in real-time using varying feedback mechanisms including haptic, visual and auditory [94]. Many of these techniques proved to be beneficial in improving mental wellbeing, displaying the need for more widespread adoption of such devices. While some feedback devices incorporated sensors, very little research has been conducted pairing physiological sensors, feedback mechanisms and AI into devices that aim to both sense and improve mental wellbeing in real-time. Overall, the feedback incorporated in a device requires careful consideration and evaluation to ensure it is effective with future devices potentially employing classification models to accurately determine when feedback should be provided.

2.6 Reflection and Challenges of Affective Technologies

2.6.1 Discussion of Existing Research

A number of systems to support mental wellbeing using apps, sensors, TUIs, biofeedback and robotics have been reviewed. A large number of apps already exist, with many aiming to improve traditional self-reporting tools and experience sampling. Apps designed to elicit PROMs provide additional convenience over traditional methods as they can be used anywhere discreetly, but self-reporting is subjective and people may fail to report [284] or be less truthful [113] when recording their mental state, showing the benefits of using objective measurements from sensors. Mobile apps reaffirm the increasing popularity of people wishing to monitor and improve their wellbeing using technological alternatives to traditional techniques.

Sensing devices are also increasing in popularity with advancements in physiological sensors resulting in cheaper and smaller devices promoting extensive use. A range of sensors have been explored measuring HR, HRV, EDA and motion. The ability to pair machine learning algorithms with sensor data presents an enormous opportunity allowing for affective states to be detected with accuracies exceeding 90% [162], [124]. Integrating sensors with classification models in a portable interface potentially enables continuous monitoring without the need to self-report. However, while AI has enormous benefits, it does present its own set of challenges, as a large amount of labelled data is first required to accurately train models.

Feedback devices aim to actively improve wellbeing in real-time using varying feedback mechanisms such as haptics, visuals and audio [94]. Haptic feedback has been used in multiple devices and often resulted in reduced stress, in particular when the feedback was personalised. Other feedback interfaces aimed to reduce stress using existing techniques such as deep breathing [186], [94], or massage therapy [56]. All these techniques proved to be beneficial, demonstrating the need for more widespread adoption of such devices. While some feedback devices incorporated sensors to monitor the impact the feedback had, very little research has been conducted pairing physiological sensors, feedback mechanisms and AI into devices that aim to both sense and improve real-world wellbeing.

2.6.2 Challenges

2.6.2.1 Privacy and Ethics

Applying therapies and translating them into digital versions is not straightforward as there are many challenges associated with mental wellbeing technologies. Privacy is a significant issue as the majority of users want to keep their mental health information private [39]. Users are more cautious regarding sharing their health data making integrating the data with established e-health systems challenging [345]. Efforts such as the General Data Protection Regulation (GDPR) in the EU and EEA have attempted to give control to citizens over their personal data by ensuring they are able to access their data and understand how it is being processed [87]. Ideally data processing should be completed locally to preserve privacy. Furthermore, care needs to be exercised regarding users' privacy with the data collected; ethical guidelines should be abided by, and users should be made aware of the data being collected and how it is being processed.

2.6.2.2 Digital Competency

An issue with some of the discussed devices is users' digital competence as elderly adults generally lack a high level of digital skills which may be required to operate these devices. However, one study [256] found elderly users preferred wearable devices over mobile phones to report emotions. Furthermore, Emoball [95] is a self-contained device and there was no evidence of digital competence affecting user interactions, showing affective TUIs can be widely adopted.

2.6.2.3 Data Collection and User Adherence

An issue with much of the existing research is very few trials collect or test using real-world data as people becoming artificially stressed in trials may not exhibit the same patterns when stressed or suffer from other wellbeing challenges in real-world situations. For real-world data collection recruiting and incentifying users to test and provide feedback on the use of such devices can be challenging, particularly regarding users' willingness to trial new affective technologies. Users will be required to trial devices to ensure their effectiveness but also to collect data enabling machine learning models to be trained. User adherence and engagement is another crucial problem as users may not immediately see the benefits of such solutions, preventing continued use. Furthermore, engaging hard to reach communities where mental wellbeing technologies could have a profound impact poses even greater challenges. Making the devices as small and portable as possible should encourage engagement as it allows them to be used anywhere [273]. The design of the devices must also be carefully considered for widespread use as they must be aesthetically pleasing to ensure the promotion of continued engagement [231].

2.6.2.4 Classification

On the diagnostic side, one of the biggest issues is affective sensing: this is inherently subjective and it may be difficult to infer through sensor data alone [299]. Machine learning models could be trained on an individual basis to allow for subjectivity to be taken into account, but this would initially require a vast amount of time and data to be collected from each user which may not be possible without first developing more accessible data collection tools. Sensing mental wellbeing not only requires accurate classification models but also accurate sensors, since if the data recorded from the sensors is not reliable, the classification from the machine learning model will not be accurate. However, when machine learning classifiers were paired with off the shelf sensors, stress was detected with similar accuracy to clinical grade sensors [209] demonstrating the potential for AI to improve accessibility to affective tools.

2.6.2.5 Portability

There are many challenges to overcome when using sensors and feedback actuators in tangible interfaces to improve mental well-being. One issue is the size of the device as it must contain sensors, a battery and feedback mechanisms which can make the device large. There are new approaches to provide feedback including Visio-Tactile feedback, that moves liquid metal drops in real-time between electrodes allowing for the feedback to be dynamic and smaller [262]. However, this is very early in development and it may not yet be possible to incorporate it into tangible devices.

2.6.2.6 Battery Life

Assuming patients are willing to use instruments used in the domain of assessing mental wellbeing, the underlying issue of battery life still needs to be addressed. Often IoT devices need to remain small and contain the necessary microcontroller and sensors leaving little room for the battery meaning it will need to be recharged regularly. A possible solution to this would be to only enable specific sensors after other actions have been performed; this means high powered sensors will not have to be continually powered but an additional step is required to collect data.

Until batteries with considerably longer battery life are developed, it will remain impractical to continually collect vast amounts of behavioural data. Instead, pragmatic solutions to optimise power consumption are necessary.

2.6.3 Summary of Current Gaps in the Research

1. Only few attempts are present in the literature that have looked at utilising multimodal real-time sensing approaches. Very often key sensors for mental wellbeing assessment are not utilised such as EDA, either because it's not available in the sensing kit or due to the fact devices with EDA sensors are often expensive [86]. Even those with multi model sensing approaches are not integrated within a full comprehensive system that also delivers meaningful feedback to the user as instant intervention. Therefore, the development of custom TUIs may improve accessibility to continuous monitoring and real-time interventions.

2. Digital markers such as physiological and human behaviour sensors and advances in edge computing present clear opportunities for real-time momentary assessment and instant intervention. However, there is an absence of large scale real-world trials to confirm the results. Many studies have used controlled experimental datasets to evaluate the models, not considering real-world performance. If affective states are to be inferred in in-situ environments then real-world labelled datasets are required to train the models. Therefore, techniques to aid the collection of real-world labelled data need to be explored.

3. The current research demonstrates two major constraints when using deep learning classifiers: a large dataset is required to first train the model and individual differences make the development of a one-size-fits-all model extremely challenging. The use of new techniques such as TL presents opportunities to reduce the impact of these limitations and to overcome data scarcity which is often associated with personal data.

4. Although the possibility of including feedback within interfaces to act as interventions has been considered there has been little focus on a single

interface that processes real-wold physiological data and provides interventions. This creates the possibility to explore new methods that automatically monitor and positively impact mental wellbeing going beyond traditional feedback mechanisms.

5. Privacy and data sharing are key concerns in the development of wellbeing devices as health data is extremely personal. Advances in edge computing are enabling instant monitoring whilst retaining privacy by not sharing data and completing all processing on-device, helping to put users in control of their wellbeing whilst also protecting their data.

Chapter 3

In the hands of users: Co-Designing Tangible Interfaces to Monitor Affective State

This chapter presents the exploration of inclusive and participatory co-design techniques and principles to engage potential users who could benefit from innovations in affective tangible technologies. In particular, individuals with cognitive impairments participated in a co-design process via a series of workshops and focus groups as their wellbeing is often diagnostically overshadowed and they can traditionally find it challenging to express their emotions [102]. The workshops helped participants explore new technologies including sensors and feedback mechanisms that can help monitor and potentially improve their mental wellbeing. The adopted co-design approach resulted in a range of effective and suitable interfaces being developed for varying ages.

3.1 Introduction

Traditional mental wellbeing assessment methods require people to be aware of their mental health status and seek help which can be challenging due to social stigma and lack of available resources [61]. The decreasing cost and increasing capabilities of sensors and edge computing is enabling new forms of interfaces which are more powerful and dynamic than traditional assessment technologies. A technological alternative that could actively monitor an individual's affective state and provide feedback could be extremely beneficial in improving accessibility to mental health tools for all [51].

Recent advances in sensors, batteries and processors have resulted in an increase of pervasive computing technologies. However, many interfaces only include a limited range of sensors such as those measuring motion and HR and infrequently record this data, resulting in a dataset that may not capture all of the wellbeing states experienced by users and is not sufficient to effectively train machine learning classifiers. Furthermore, the vast majority of current interfaces that monitor wellbeing do not include methods that can serve as intervention to promote the regulation of negative emotions [114].

TUIs enable people to interact with digital data through physical objects and are ideal to embed all of the necessary sensors and processing power [317]. Tangible manipulation presents an opportunity to develop novel devices introduced as tangible fidgeting interfaces that go beyond traditional devices to enable unique interaction methods and encourage continued engagement [136]. A successful TUI is envisioned that will enable affective state to be automatically monitored using multimodal sensors embedded within the device. The possibility of a device to accurately monitor real-world wellbeing using sensor data, further enables the possibility of applying digital feedback to act as interventions aiming to automatically and gradually regulate emotions and improve wellbeing in real-time.

A person diagnosed with an intellectual disability [274] shows deficits in intellectual functioning such as reasoning or problem solving with an IQ score often less than 70 and deficits and impairments in adaptive functioning such as communication or social skills, all during the developmental period before the age of 18 [11]. Intellectual disabilities can be divided into four levels: mild, moderate, severe and profound where for each level the person requires more support [196]. Individuals with intellectual disabilities often experience mental wellbeing challenges but these are frequently overlooked and attributed erroneously to their disability (diagnostic overshadowing) or classed as challenging behaviour [102]. Diagnostic overshadowing often results in a significant impact on an individual's likelihood of engaging with mental healthcare systems. Furthermore, many individuals with intellectual disabilities find it challenging to express the emotions they experience [3]. Individuals with intellectual disabilities face additional challenges in expressing their emotions, correctly interpreting social situations and predicting the behavioural consequences of specific actions [46], [263]. This research seeks to address these challenges using a co-design approach.

To ensure the proposed TUIs are effective and usable by individuals who experience mental wellbeing challenges, including those with intellectual disabilities, it is imperative to design the interfaces with end users. An iterative co-design process has been proposed that has been adapted to enable the designing of interfaces and exploration of different sensing and feedback mechanisms with participants from the Nottingham Interactive Community for Education Research (NICER) group at Oak Field school, Nottingham, UK. The group is formed of adults with a range of intellectual disabilities who have a wide range of experience in evaluating enabling technologies, including virtual environments and serious games. This approach hopes to alleviate communication challenges faced by the target group and gather feedback that results in the development of interfaces that can effectively monitor real-world affective state.

The aim of the iterative co-design process was not to achieve one design with specific sensors, instead a range of different devices was expected to be proposed to suit the needs of all potential users including those who may experience a wide range of limitations. The co-design process aimed to investigate the following research questions 1) How can the co-design process be adapted to best suit the needs of those with intellectual disabilities? 2) What are the optimal design guidelines for prototyping affective tangible interfaces? 3) Which sensors would be most beneficial for user interaction? 4) How do users believe on-device feedback could help improve their mental wellbeing?

3.2 Background of Co-design

Co-design is closely connected with participatory design and is the methodology for actively engaging people directly involved in an issue, place or process in its design, allowing them to make a meaningful contribution to the design process [49], [32], [134]. Co-design enables the reduction of the gap in knowledge between end users and researchers, allowing non-designers to become equal members of the design team, ensuring designer subjectivity is removed and the technologies developed are suitable for the target population [325], [269]. During the process, design tools are used to empower all of the participants to facilitate a 'joint inquiry' where 'problem and solution co-evolve' [295]. Co-design brings many benefits to the design of the project by helping the researcher better understand the challenges faced by users and any potential solutions [296], [287].

Co-designing helps solve real-world problems by bringing together people from different backgrounds into the design process, resulting in more inclusive solutions. However, to work most effectively it is important to select appropriate methods and ways of working which need to match the project being designed and the potential users' capabilities and limitations. Co-design methods help make things that are normally unobservable through traditional interviews and focus groups available as resources for design [326], [82], [327], this can be achieved by:

- using visual, creative methods [266].
- physically making things helping people to explore, verbalise, remember and imagine [269].
- creating and telling stories helping to put things into context and providing a central way of sharing and communicating. Story sharing can be visual, verbal or include role play. [266].

Co-design can be used to promote the inclusion of people living with disabilities when designing new solutions by including their personal experiences, making them more likely to take ownership of the final outcome [272]. People with intellectual disabilities may face barriers such as communication challenges when being involved in the co-design of new assistive technologies, resulting in co-design techniques needing to be modified to fit with participants' abilities [34]. To help reduce the challenges faced, a set of guidelines have previously been devised for co-designing with people who have autism spectrum disorders [93].

Co-design has previously been used to engage individuals with cognitive disabilities in successfully designing a picture-based remote communication system, helping the participants move from merely being passive onlookers to active participants [71]. Similarly, co-design workshops have engaged people with assisted living needs to develop technologies and services for new care solutions [342], engaged autistic children in co-designing technologies [292], researched accessible apps and games [15] and helped people with complex communication needs express themselves through art therapy [181]. Furthermore, co-design has been used to develop mobile applications with adults who have intellectual disabilities, highlighting that prototypes were required to deepen user engagement [286]. Overall, while involving users with intellectual disabilities in the design process produces additional challenges such as additional ethical considerations [293], it is imperative to ensure the solutions developed meet the needs of the potential users.

3.3 Participatory Iterative Design Methodology

This instantiation of an iterative design cycle took over one year in total. Each stage of the co-design process was conducted with the same primary researcher and an experienced facilitator with many years experience in running co-design workshops. All participants were members of the NICER group with varying disabilities including Williams syndrome, Down syndrome and Autism but no participants had significant motor skill impairments that would impact their participation. Their role in this process was not as research subjects, they were instead involved in identifying design opportunities relevant to their needs.

Co-designing with people with disabilities presents additional challenges - such as issues regarding the nature of their cognitive disability, including communicational and memory issues, which may challenge their full participation. However, it is imperative for the voice of end users to be heard and many of the challenges can be overcome by developing a co-design methodology catering for the requirements of people with intellectual disabilities. The early integration of people with intellectual disabilities into the design processes aimed to prioritise their design decisions and needs. These insights serve as guides to a joint inquiry that seeks to address the challenges of developing TUIs to monitor affective state.

The co-design process used in the project is depicted in Figure 3.1 adapted

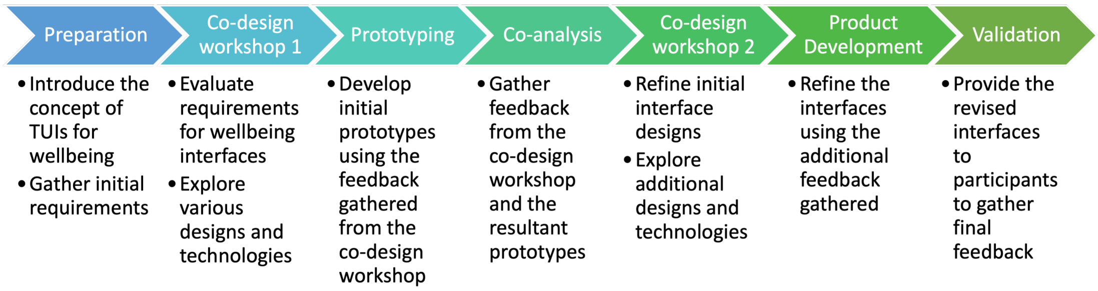


Figure 3.1: Stages of iterative co-design process.

from [330], [331] where people with intellectual disabilities engaged in a partnership with researchers. The methodology alternates between focus groups to gather feedback during the preparation, validation and evaluation phases and interactive co-design workshops completing the fieldwork and ideation phases. A novel aspect of this process is that it enabled feedback to be gathered at each stage of the development process including design and prototype development. Overall, the co-design process was conducted in conjunction with an advisory panel of adults with intellectual disabilities, focusing on methodological adaptations and special supports developed to facilitate and ensure their participation. The co-design sessions were recorded for future analysis as granted by Nottingham Trent University, ethics application 18/19-43V2.

3.3.1 Preparation Phase

At the beginning of the co-design process an introductory preparation phase was completed to agree on the scope and aims and objectives of the project. Members of NICER along with teachers of young students with moderate to severe intellectual disabilities acted as an advisory panel with a researcher, experienced facilitator and education specialist leading the session. The focus group was conducted over 1 hour and notes were recorded for future analysis.

An accessible introduction was completed using a presentation to introduce participants to the concept of TUIs and the possibility for them to automatically monitor affective state. This was completed by showing examples of existing interfaces such as Emoball [95] and Mood TUI [273] to help participants develop a concrete understanding of TUIs and allow them to gain greater knowledge of the devices to be developed. Explanations were provided of hard to understand concepts including discussions on emotions and the technologies themselves.

After the project had been introduced participants were asked whether it was of interest to them, and whether they would like to get involved in the subsequent co-design workshops and they agreed and volunteered their involvement. The NICER group is a significant stakeholder in this research, and they developed high expectations of TUIs that could have a positive impact for the intellectually disabled community.

3.3.2 Co-design Workshop 1

The first co-design workshop aimed to explore various designs, technologies and requirements for affective TUIs. The workshop was conducted over 4 hours at Nottingham Trent University to strengthen the participants' roles as experts and posit them as co-researchers within a university setting and was video and audio recorded for future analysis. The workshop comprised of six participants; four males and two females who have previously been involved with multiple research projects and are experienced co-designers. Table 3.1 shows the Number (N) of participants and their characteristics including those who have Williams Syndrome (WS), Down Syndrome (DS) and Autism along with information on their gender and level of intellectual disability as defined by their condition and previous evaluation by professional carers and teachers.

A common issue with the development of mental wellbeing technological solutions is the lack of ethical considerations [265]. To ensure the co-design workshops were inclusive for participants with intellectual disabilities and caused no harm, all discussions were short and a range of interactive tasks were designed to increase engagement. The following five design affordances used within the co-design workshops were designed to support participants' decisions, compensate for their disabilities and act as conversational instruments based on established methods [291] and previous experience [45], [42].

Table 3.1: Co-design workshop 1 participant characteristics (age, gender disability) for total Number of participants (N), with Williams Syndrome (WS), Down Syndrome (DS) and Autism.

		Total (*N=6*)	WS (*N=1*)	DS (*N=4*)	Autism (*N=1*)
Mean age (range)		36 (18-47)	39	39.75 (30-47)	18
Gender *N* (%)	Men Women	4 (66.7) 2 (33.3)	1 (100) 0 (0)	2 (50) 2 (50)	1 (100) 0 (0)
Level of intellectual disability *N* (%)	Moderate Severe	1 (16.67) 5 (83.33)	0 (0) 1 (100)	1 (25) 3 (75)	0 (0) 1 (100)

3.3.2.1 Introduction and Demonstration

When conducting co-design workshops with participants who have intellectual disabilities it is imperative to ensure all participants fully understand the goal of the workshop to improve communication. To ensure this, an experienced facilitator introduced the session by clearly explaining the concept of affective tangible interfaces.

Challenges can arise from communicating with participants and difficulties interpreting non-verbal interactions [126], [224]. Therefore, in the co-design processes the adage "show me don't tell me" [49] is often used, resulting in previously developed prototypes such as the cube shown in Figure 7.1 being demonstrated to the participants. Concrete prototypes, create opportunities for participants with intellectual disabilities to interact directly with the interfaces and understand the feasibility of developing new interfaces [40]. The prototypes embedded a range of sensors including 9 Degree of Freedom Inertial Measurement Units (9-DOF IMU) to measure motion, Force Sensitive Resistors (FSR) to measure touch and HR and EDA sensors to measure physiological changes.

3.3.2.2 Storyboarding and Drawing

The design of the interfaces was then explored. Previous work provides many useful strategies for engaging individuals with intellectual disabilities in the design process such as storyboards and pictures, and avoiding open-ended ques-

Figure 3.2: Original mental wellbeing interface prototype.

tions [37], [97], [255], [288], [355]. To help compensate for the participants' intellectual disabilities, real-time storyboarding was completed where prompts were presented to participants to expand upon, promoting communication. During storyboarding, participants were able to discuss their opinions on the existing interfaces previously demonstrated and share their ideas for new interfaces.

In co-design, methods are used to help participants 'say, do and make' [268]. This helps us deepen our engagement with people and strengthens the insights we are able to gather. Using this approach participants were invited to draw their own interfaces using pen and paper to help promote ideas for new interfaces. This enabled the participants to creatively express their design ideas without the need to verbalise, which those with intellectual disabilities can find challenging.

3.3.2.3 Prioritising Requirement Cards

The potential features of tangible interfaces were also explored as it is imperative to understand what features users most require to ensure successful devices are developed. A card based approach was used that enabled participants to prioritise the features they believed were most required. This approach was based on the generative research approach [267] to combine participatory exercises with verbal discussion during the creative idea generation phase. Similar card based approaches have previously been used due to their accessibility, familiarity and tactile nature which can help promote communication [33]. Six cards were provided to each participant stating a specific requirement for tangible interfaces including: *ease of use*, *makes me feel better*, *design*, *battery life*, *physical size* and *understands how I am feeling*. Each of the six requirements were explained to the participants by the researcher and experienced facilitator to ensure they fully understood the meaning of the requirements and their role in prioritising the requirement cards.

3.3.2.4 Real-time 3D Printing

The process of showing participants with an intellectual disability how their design decisions have a direct real-world consequence in a rapid and concrete way was developed in an earlier study and replicated here [44]. 3D modelling software was used to demonstrate how the interfaces can be designed and printed to make concrete the relationship between the participants' decisions and the tangible interfaces produced. A majority vote was conducted to decide on the shape to be printed and as the workshops were conducted over several hours there was sufficient time to design and 3D print a small interface, providing opportunity for participants to provide reflection on the design. Creating the interfaces during the workshops resulted in a deeper and more practical understanding about the participants' experiences [268].

3.3.2.5 Interactive Electronics

When exploring new technological solutions, providing demonstrations is necessary to ensure all participants understand the functionality and how the technology can be used, thereby improving confidence and communication [41]. During the session a range of non-invasive, easy to use sensors that could be used in real-world environments [6] were explored through interactive demonstrations. This was designed to increase engagement and ensure participants understood the functionality of the electronics by allowing them to experience the different capabilities offered by each sensor. All electronics were made simple to operate with the electrical circuits pre-built, as used in previous co-design studies [20], to ensure all participants would be able to fully participate [133].

A HR sensor was first explored where participants were able to place their finger on the sensor and lights would flash at the same rate as their pulse. An

EDA sensor was also explored as it functions in a similar way to the HR sensor with participants having to place their fingers on the sensor. An FSR was demonstrated next, where as participants pressed harder on the sensor it caused a haptic motor to vibrate. Finally, a 9-DOF IMU was demonstrated inside a ball; as participants shook the ball it would vibrate. Overall, this method of exploring the sensors promoted participants' understanding and enabled them to experience how the sensors will be used in future interfaces.

Varying forms of feedback acting as real-time interventions were also explored giving participants time to reflect and express their feedback between demonstrations. Visual feedback continued to be explored following the use of multicoloured LEDs to demonstrate the HR sensor. Participants were shown multiple examples including one device where different colours represented different emotions. Auditory feedback was also demonstrated where a speaker was used to play calming sounds from nature. Haptic feedback was the last intervention explored; four different feedback patterns were demonstrated to each of the participants who held the vibration motor to experience the different sensations. The exploration of interventions through interactive sessions enabled all participants to experience each of the feedback mechanisms, allowing them to provide personal insights and share their immediate impressions.

3.3.3 Prototyping

Using the feedback from the co-design workshop, initial high-fidelity prototypes were developed using the designs, sensors and feedback mechanisms suggested. This rapid prototype development was conducted to enable participants to physically experience their design suggestions from the workshop, promoting further discussion and the continued refinement of the interfaces.

3.3.4 Co-analysis

During the co-analysis phase, focus groups were held where the participants who took part in the co-design workshop, along with other members of the NICER group and teachers of young students with intellectual disabilities, were provided with the opportunity to give feedback on the workshop and the resultant ideas.

Typically, the focus groups were shorter in nature than the co-design session allowing for reflection on the outcomes of the workshops. During the co-analysis notes were recorded for future analysis.

The various activities conducted within the workshop were discussed to analyse what activities participants enjoyed and how they though the workshops could be improved, allowing for adjustments to be made to the co-design process where required. The group facilitator asked the NICER group members who attended the co-design workshop to present the main activities and outcomes of the workshop and their implications. Volunteers presented their memories, experiences and design preferences and this again was all recorded using an accessible storyboard format. Implications for adjustment of co-design techniques and plans for follow up activities at the next co-design workshop were also discussed.

The initial prototypes based on the feedback gathered from the co-design workshop were also demonstrated to participants. This allowed participants to provide feedback and describe each functional prototype to the rest of the group, with the experienced facilitator and researcher eliciting further feedback on favourite designs, future design iterations and considerations.

3.3.5 Co-design Workshop 2

The second co-design workshop advanced upon the findings from the first workshop and co-analysis aiming to refine the developed interfaces. The workshop was conducted over 3 hours and was audio recorded for future analysis. The workshop comprised of 8 participants; 5 males and 3 females who again are experienced co-designers. 5 participants previously participated in the first co-design workshop, these included 1 with Williams syndrome and 4 with Down syndrome. Table 3.2 shows the characteristics of the participants including those who have Williams Syndrome (WS), Down Syndrome (DS), along with information on their gender and level of intellectual disability.

During the co-design workshop the affordances previously developed were used. The session was introduced and the previously developed interfaces were demonstrated to the participants, ensuring all participants were familiar with the interfaces and understood the purpose of the co-design session. After participants

Table 3.2: Co-design workshop 2 participant characteristics including Number of participants (N), with Williams Syndrome (WS), Down Syndrome (DS) and Autism.

		Total (*N=8*)	WS (*N=2*)	DS (*N=4*)	Other (*N=2*)
Mean age (Range)		36.5 (27-47)	38.5 (38-39)	39.75 (30-47)	28 (27-29)
Gender *N* (%)	Men Women	5 (62.5) 3 (37.5)	2 (100) 0 (0)	2 (50) 2 (50)	1 (50) 1 (50)
Level of intellectual disability *N* (%)	Moderate Severe	1 (12.5) 7 (87.5)	0 (0) 2 (100)	1 (25) 3 (75)	0 (0) 2 (100)

understood the aim of the session was to refine the existing prototypes, storyboarding was again utilised to help participants effectively communicate their new ideas for the interfaces. However, following feedback from the co-analysis and the challenges encountered during the first workshop participants did not draw their own interfaces. Instead, 3D modelling was performed and participants experienced 3D printing a new interface during the workshop to help them understand the prototyping process and ensure they understood the impact their design decisions have on the developed interfaces.

The electronics within the interfaces were also explored with participants experiencing the same sensors as previously explored to ensure the suitability of the sensors and help gather additional feedback from new participants who hadn't previously experienced the electronics. Finally, potential features were not prioritised using cards as conducted in the initial workshop due to this workshop focusing on refining the ideas already produced. Instead, further discussions were held enabling participants to express their opinions on the required functionality and how the prototypes could be improved.

3.3.6 Product Development

Using the feedback from the second co-design workshop, the initial prototypes were refined. Multiple refined interfaces were developed over several weeks, each considering the feedback gathered throughout the iterative co-design process.

These final products allow participants to experience how their design decisions impacted the development of solutions relevant to themselves.

3.3.7 Validation

The final aspect of the iterative co-design process was to evaluate the developed interfaces. Each of the developed interfaces were demonstrated to members of the NICER group and teachers of young students with intellectual disabilities where notes were taken for future analysis. The participants had the opportunity to experience the different interfaces developed throughout the co-design process and examine how their feedback helped influence the design and functionality of the devices. Participants also had the opportunity to provide their final feedback on all concepts generated, including the design and technologies within the interfaces in addition to the co-design process itself and how it was adapted to promote communication and idea generation.

Finally, volunteers were sought to take the TUIs home for longer term testing and labelling. At the end of these label collection phases, feedback on the usability, durability and performance of the devices in home settings was also sought.

3.4 Results

3.4.1 Thematic Analysis

Handwritten notes, video recordings and audio recordings of the co-analysis and co-design workshops have been analysed using thematic analysis [36]. During thematic analysis, codes are used to describe specific topics from the recordings and then themes are created from the codes. When conducting thematic analysis of the data, whenever a possible feature to include or exclude was discussed, it was coded. This approach highlighted all of the recommendations suggested by participants during the co-design workshop and the justification of each feature.

After coding the sessions, similar codes were grouped together to create initial themes. Each theme contains feature suggestions or restrictions, answering the research questions used to create the codes. Each code was then checked within the themes to ensure its suitability, resulting in minor changes to the themes. The thematic analysis process resulted in seven themes being created from the codes which were all consistent and represented the feedback gathered. For each of the seven themes discussed, relevant excerpts from the recordings are provided to highlight the opinions of potential users experiencing mental wellbeing challenges. The complete list of recommendations organised by themes is shown in Table 3.4.1.7.

3.4.1.1 Design and Personalisation

When considering the design of the interfaces size was a key factor, it was stated the interfaces should be *"not too big and not too small"*. The size of the existing cube prototype was liked by participants with them stating *"like that size"* when referring to the size of new interfaces. However, when discussing compromises with size all participants agreed they would prefer slightly larger devices with additional sensors, stating *"a big device with everything... would be really handy"*. Participants also wanted the ability to use it anywhere so a rechargeable battery is necessary. Participants discussed numerous usage scenarios for the portable interfaces including using them as work, college and at home. Participants envisioned carrying the device with them and possibly placing it on their desk allowing them to use it whenever they felt necessary to help them understand their affective state or help them relax.

When exploring the existing prototypes opinions varied as to whether the hard 3D printed shapes or the soft interfaces were preferred. Most participants suggested the softer devices could represent toys and be developed for younger children while the 3D printed interfaces could be reserved for older children and adults. The separation of devices for different age ranges had not previously been considered but participants believed this would be beneficial to reach the most people who may benefit from the technology. Drawing on their own past experiences and examining the developed prototypes helped participants suggest these new ideas that they believed would be beneficial for children to promote inclusivity.

The colour and personalisation of devices was repeatably mentioned by participants and in particular it was stated the devices should *"not be black and white"* for children and participants would like the ability for the devices to *"change different colours"*. When asked whether everyone could use a similar device or whether personalised devices would be required, participants suggested the idea of a base interface being developed that can then be customised with cases to make the device more personal.

3.4.1.2 User Engagement

When considering potential users of the device, children were suggested multiple times throughout the workshops. It was agreed that younger children aged around 5-8 (middle childhood) would be engaged by softer interfaces stating "*these are for the kids"* . It was also stated that the larger soft devices may be more appropriate for younger children with less parts to break or chew. The 3D printed interfaces were considered less appropriate for children but would be more engaging for older children and adults to use during everyday life due to their inconspicuous design.

When drawing future interfaces, a novel method to personalise the interfaces by attaching extensions that contain additional sensors or feedback was devised to increase engagement. The addition of extensions would enable the devices to adapt to the user enabling the most beneficial sensors and feedback to be included on an individual basis.

3.4.1.3 Sensor Inputs

When the device inputs were explored, participants were successfully able to use all of the tested sensors. Participants found the IMU and FSR sensors the easiest and most natural to use. However, not all sensors would be appropriate for children, such as the HR sensor, as children would not be able to keep their finger in continuous contact with the sensor for the device to accurately measure physiological changes. It was suggested that physiological sensors should be reserved for the 3D printed interfaces and not the soft interfaces, where children in particular may continuously move their hands preventing accurate readings from being recorded.

The use of accelerometers to measure motion was well received with participants finding it useful for the device to recognise how it is being handled such as being bounced or thrown. Accelerometers could be easily embedded within all interfaces, with participants believing the manner in which the devices will be interacted with will be different depending on the user's state of wellbeing. Similarly, FSRs to measure touch could be embedded within both 3D printed and soft interfaces. The ability to measure touch is useful as participants believed stroking the soft interfaces was relaxing as it simulated stroking a pet. Participants enjoyed interacting with touch to activate the feedback, such as pressing hard to enable the visual and haptic feedback as they found this method of interaction intuitive. The way in which the device was touched was also suggested as a mechanism to indicate wellbeing, with users potentially squeezing the device harder when angry. Overall, the touch sensors and accelerometers were preferred by participants as they were simple to interact with and can be embedded within all interfaces for all potential users.

The ability for the interfaces to record messages was a popular request for the soft interfaces. Participants stated they would like to *"talk to it about how you feel"* and they would *"want it to record what is being said ... and play it back"*. Participants believed the soft toy-like interfaces could act as friends for children where they could express their emotions. The inclusion of buttons within the interfaces was also proposed, with participants liking the buttons to fidget with to potentially improve wellbeing and to initiate feedback. A keyboard input was also suggested, although inclusion of this feature would result in the keys performing the same function as buttons. The majority of the inputs suggested can be implemented using force sensors to measure touch, stroking and squeezing, IMU to measure shaking, HR and EDA sensors to measure physiological changes and buttons for fidgeting.

3.4.1.4 On-device Feedback

When considering outputs, the ability for the devices to make sounds was frequently mentioned. Participants liked the audio feedback playing calming sounds but it was suggested that alternative sounds could also be played to improve mood. Participants believed the ability for interfaces to play calming sounds, music, stories or recorded messages may greatly improve wellbeing. Guided relaxation was also explained to participants as a potential method to assist relaxation, which participants preferred along with recorded messages. However, a common criticism with the auditory feedback was the quality of the reproduced sound. This could be improved by using a larger speaker, although this would increase the physical size of the device to larger than what participants believed would be comfortable.

Visual feedback was suggested multiple times throughout the workshops, suggestions included *"letting people see different shapes and colours"*, *"make it flash either once or twice"*, *"different colours like traffic lights"*, *"lights up bright colours"* and *"disco lights"*. When participants designed their own interfaces, the majority included lights activated through either buttons or touch. Visual feedback was the most popular choice for inclusion in the children's interfaces where it was suggested that lights should be included throughout the interfaces to make them easy to see with *"different shapes and colours of lights"*. Participants found the inclusion of lights to be vital in making the interfaces more engaging as well as helping them to feel relaxed. Varying combinations of lights were trialled, with participants preferring light matrices enabling different patterns to be displayed in addition to varying colours. However, when participants were shown an example where different colours represented different emotions, some were not able to understand the concept. This suggests that visual feedback can be used as a distraction technique aiming to improve wellbeing but not to convey information to those with an intellectual disability.

The possibility of embedding a screen within the interfaces was suggested to show pictures and *"sometimes show happy, sometimes show sad faces"* dependent on the participant's inferred state of wellbeing. While participants suggested adding a screen, they also suggested that directly showing measures of stress through the feedback may induce additional stress, as has been previously found [197]. Therefore, to ensure no additional stress is caused, visual feedback should only be used to display varying patterns to aid relaxation.

All participants enjoyed experiencing haptic feedback as shown in Figure 3.3

Figure 3.3: Participants exploring different haptic patterns during a co-design workshop.

and preferred it in comparison to the other feedback mechanisms. Upon first experiencing the haptic feedback, participants stated *"I love it"* and *"it's amazing"*. When testing different haptic patterns the majority of the participants preferred pattern 1, which involved long subtle consistent vibrations that gradually reduced compared with the second most popular choice, pattern 4, which involved short sharp fast vibrations that also gradually reduced force stating *"It's quite soft"* and *"that's better"*. The least favoured patterns included pattern 2 which gradually reduced the force of the vibration in 20% steps similar to pattern 1 but less gradual and pattern 3 which alternated between maximum force and slightly reduced force. Participants didn't like the stepped nature of pattern 2 and also didn't like the consistent high force of pattern 3 showing these sharp, powerful vibration patterns should not be used to relax users. It is not unexpected that pattern 1 was most preferred due to its slow gradually reducing, relaxing nature, however it was surprising that pattern 4 was the second favoured pattern due to its harsh nature that would not usually be considered relaxing as shown in Figure 3.4. When comparing haptic feedback patterns, one participant stated *"all the rest are harder number four is soft"*, showing that it may be necessary to personalise haptic patterns on an individual basis. Participants also suggested the haptic feedback patterns should remain slow, potentially slower than the user's heartbeat, to ensure they are relaxing. A possible challenge with haptic feedback is users becoming accustomed to the sensation, although the continuous variation of patterns 1 and 4 helped to alleviate this issue with users believing it would help them become more aware of their emotions and help them remain calm.

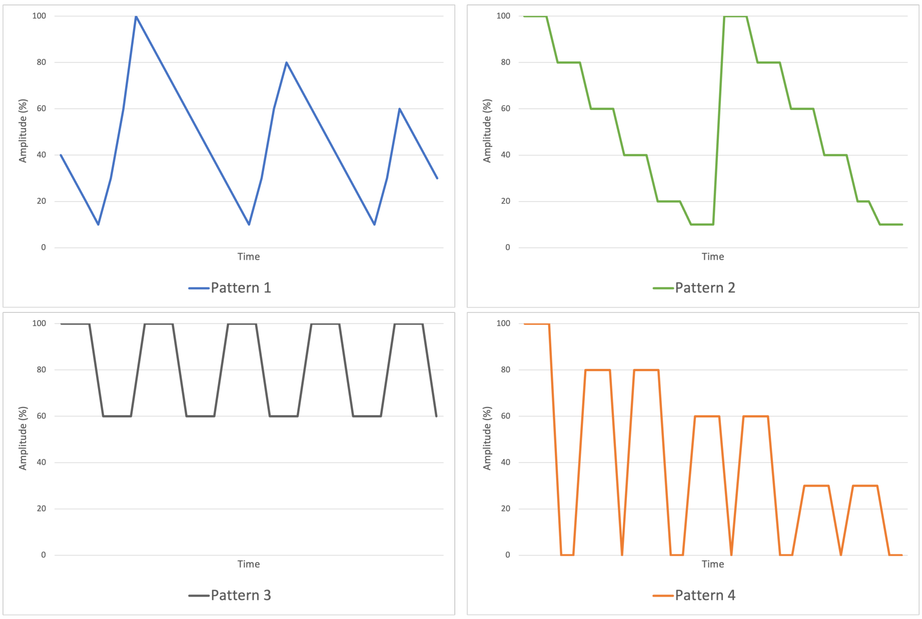


Figure 3.4: Percentage of vibration amplitude for the four tested vibration patterns.

Finally, participants made suggestions regarding the ability of the devices to move and change shape. Participants requested the ability for the interfaces to change shape, stretch or *"become rounder"*. Flexible electronics may enable the interfaces to move in this way using motors, although changing shape remains challenging for 3D printed devices.

Overall, participants suggested numerous outputs for the design of future devices. Based upon user feedback, haptic feedback is a key intervention that should continue to be explored as all participants found it relaxing and preferred similar vibration patterns. Visual, haptic and auditory feedback can all be embedded within future devices, although shape shifting may be more challenging and the addition of a screen would require careful consideration to ensure it does not induce further stress. Overall, participants believed by including these feedback mechanisms future interfaces could improve their mental wellbeing in real-time.

3.4.1.5 Design Requirements

During the first workshop participants also explored the requirements of affective tangible interfaces. Table 3.3 shows the order in which five participants prioritised the requirements as one participant was unable to complete the task.

Table 3.3: 5 Participants' tangible interfaces priorities as ordered during the co-design workshop from highest to lowest priority.

Participant 1	Participant 2	Participant 3	Participant 4	Participant 5
Ease of use	*Battery life*	*Makes me feel better*	*Size*	*Battery life*
Makes me feel better	*Makes me feel better*	*Understands how I am feeling*	*Makes me feel better*	*Ease of use*
Design	*Ease of use*	*Ease of use*	*Understands how I am feeling*	*Size*
Battery life	*Design*	*Size*	*Battery life*	*Design*
Size	*Understands how I am feeling*	*Design*	*Ease of use*	*Understands how I am feeling*
Understands how I am feeling	*Size*	*Battery life*	*Design*	*Makes me feel better*

Each requirement, was given a score dependent on the order in which each participant placed the requirement, where the highest priority was given a score of six and the lowest one. The most prioritised requirement was *makes me feel better* (22), closely followed by *ease of use* (21), then *battery life* (19), *size* (16), *understands how I feel* (14) and finally *design* (13). This shows that the participants all value the feedback the device could provide to make them feel better as the highest priority. However, this would first require the device to understand how the individual is feeling which was second least prioritised feature, possibly showing a lack of understanding of the *"understands how I am feeling"* requirement. This is not unsurprising as this requirement is the most complex and requires the understanding that a computer model is capable of interpreting affective state. *Ease of use* was the second highest rated priority showing it to be highly valued amongst all participants. *Battery life* and *size* followed, although this greatly varies between participants, with some participants rating them as the highest priority and others rating them as the lowest priority. The lowest overall priority was *design* which was unexpected as participants enjoyed exploring the different prototypes suggesting they see *makes me feel better* as the overall design goal.

3.4.1.6 Design Limitations

During the workshops numerous limitations were discussed but all participants considered it highly important that the interfaces did not appear as a medical device in order to reduce stigma. When a participant previously used medical sensors they stated *"I felt awful, I was panicking...the first time I thought I'm not doing this"*. This makes it vital that any physiological sensors and feedback mechanisms within the interfaces are non-invasive, easy to use and inconspicuous, as to not induce additional stress. As the sensors explored during the workshops were all small and unobtrusive, participants believed they were ideal to monitor real-world affective state. By developing wellbeing interfaces for the general population, as well as for those experiencing mental health challenges, it will reduce the associated stigma by ensuring the devices are suitable for all.

When exploring haptic feedback, all participants liked the varying haptic patterns but when initially experiencing the full powered haptic motor, it was *"too much"* for one participant, while another did not like the noise when the motor vibrated against the table. This demonstrates that while haptic feedback was enjoyed and should be explored, the haptic patterns played within the devices can have a major impact on enjoyment and wellbeing. On the other hand, when the haptic motor was embedded within a large soft toy one participant could not feel the motor, showing a careful balance is required to ensure the haptic feedback can be felt within the interfaces but is not overpowering.

Finally, participants stated they would like to have all of the sensors and feedback within one device, however this would drastically decrease battery life and increase the size and cost of the interface. The soft interfaces contain ample space to embed a variety of sensors and feedback. However, the 3D printed interfaces are very limited in space especially if they are designed to be small enough to use in the real-world, resulting in only those sensors and feedback preferred by the majority of participants being included.

3.4.1.7 Cognitive Barriers

During the workshops cognitive barriers was a theme that appeared throughout, mainly through lack of communication. Frequently, participants would require prompting when discussing specific topics, for example when explaining which sensors would be most appropriate for use. There was a large variation in the communication skills within the group, with some participants elaborating on their feedback in great detail and others who frequently replied with one-word answers, simply nod or shake their head or always agree with the other participants. These are challenging issues to overcome as some participants may not feel comfortable in expressing their opinions in groups. However, the involvement of an experienced facilitator who understood the commonly occurring communication issues for this group, the use of Makaton symbols where appropriate along with the interactive activities during the co-design workshops, helped improve communication with all participants.

Providing initial prototypes for participants to explore aided their understanding of potential uses for the devices, helping them to communicate new ideas and features and the interactive sessions were all received extremely positively. These sessions helped participants gain a clearer understanding of the interfaces and the technologies used within them, helping many to better communicate their ideas and hence provide qualitative data to help answer the research questions.

Drawing new interfaces helped some participants creatively express their ideas that they were not able to verbally communicate such as the sensors and feedback they believed were most important to include, although other participants struggled to draw an example device. Allowing participants to express themselves non-verbally through 3D printing during the workshops helped demonstrate how the interfaces are developed and encouraged all participants to consider how different shaped interfaces could be used. This process helped participants make a concrete connection between their design decisions (their drawings) and how this had a direct impact on the outcome (3D printed interface) [44]. Ranking features of the interfaces was slightly more challenging for some participants, although after additional guidance the majority of participants successfully prioritised the features. This helped them express what functionality they believed was most important without needing to verbally communicate. Finally, exploring the sensors and feedback was enjoyed by all and the analysis showed it greatly improved attention and engagement in addition to aiding the understanding of each technology. Demonstrating how the electronics work aided participants' understanding of how the devices are developed and helped them make realistic suggestions regarding the design of future tangible interfaces.

Table 3.4. Themes and guidelines devised from thematic analysis

Theme	Guideline	Description
Design and Personalisation	Battery life	The interfaces must include a rechargeable battery that can power the device throughout the day where it can be used in real-world environments but must also remain small enough to fit within the interfaces.
	Colour	The interfaces should be developed in a variety of colours enabling them to feel personal.
	Shape	A variety of shapes including 3D printed interfaces and softer interfaces such as children's toys should be developed to cater for a wide variety of users experiencing differing challenges.

Continued on next page

Table 3.4. Themes and guidelines devised from thematic analysis

Theme	Guideline	Description
User Engagement	Children	Soft children's toys should be developed that embed the necessary child friendly electronics and interventions such as visual and haptic feedback to improve engagement.
	Extensions	A fundamental interface could be developed that can then be expanded upon to customise the sensors and feedback on an individual basis.
Sensor Inputs	Physiological	Sensors to measure physiological changes such as HR and EDA should be utilised to help monitor affective state but their use may not be suitable in soft children's interfaces.
	Motion	The use of IMUs to measure motion can easily be embedded within all interfaces and may show how device interactions differ with wellbeing state.
	Touch	The ability to recognise where and how hard the device is touched may help infer wellbeing, with stroking possibly providing relaxation.
	Buttons	Buttons to promote fidgeting or to activate feedback should be included to improve individuals' wellbeing.
	Microphone	It may be possible to infer affect via voice although this imposes privacy concerns.

Continued on next page

Table 3.4. Themes and guidelines devised from thematic analysis

Theme	Guideline	Description
Feedback	Auditory	Calming sounds, guided relaxation, stories or voice recordings could be played to aid relaxation.
	Haptic	Calming haptic feedback patterns can be played when a user is stressed to help improve wellbeing.
	Visual	Lights can be displayed in various shapes and colours to convey information or act as a distraction to calm users.
	Tactile Interactions	The possibility for interfaces to physically move or change shape could provide therapeutic tactile feedback.
Limitations	Size	The interfaces need to remain small enough to be portable while also embedding the required sensors and feedback. Individuals may have their own preferences concerning which sensors and types of feedback function most effectively, resulting in personalised devices.
	Stigma	It is vital the interfaces do not resemble medical devices to make them ubiquitous and reduce stigma as the interfaces should benefit anyone experiencing mental wellbeing challenges.

Continued on next page

Table 3.4. Themes and guidelines devised from thematic analysis

Theme	Guideline	Description
	Haptic sensation	Haptic feedback should be felt within the interfaces but it must not be overpowering to ensure it has a positive impact on wellbeing.
Cognitive barriers	Prototypes	The development of initial prototypes aided the understanding of the functionality required within the interfaces, helping to improve communication.
	Interactivity	Interactive sessions improved engagement and enabled greater communication regarding the design, sensors and feedback for the wellbeing interfaces.

3.4.2 Product Development

A number of prototypes were developed and categorised into soft toys for young children aged 5-8 (middle childhood) as this is when children develop relevant social, emotional and cognitive skills [118] [260] and 3D printed interfaces for older children (8+) and adults. Age is an important factor to consider when developing tangible interfaces to ensure they are accessible and engaging, helping to reduce stigma with affective tools. Privacy was a key consideration when developing the interfaces. Therefore, the interfaces do not communicate with any external devices as all data processing and storage is local. Additionally, microphones were not included within the interfaces to capture audio data as the continuous recording of voice to infer wellbeing is highly sensitive and may raise privacy concerns [172]. Speakers were also dismissed due to the requirement for large speakers to play better quality audio.

Initially, the soft children's prototypes were developed. Learning from the feedback gathered, physiological sensors were omitted from these interfaces, instead only touch and motion interactions would be measured. An FSR sensor was initially explored to measure touch, however the rigid shape of the sensor and limited contact space reduced its potential utility. Conductive fabrics were then explored as an alternative capacitive sensor to measure resistance. Conductive fabric can be shaped to cover the entirety of the interface enabling all touch interactions to be detected. Using conductive fabric as a capacitive sensor enables the location and pressure of interactions to be measured with the same accuracy as an FSR sensor, while additionally enabling all of the interface's surface to be monitored. To measure motion, a 9-DOF IMU continued to be used as its small size and inclusion of an accelerometer, gyroscope and compass provides a large sample of data to monitor interactions.

Next, the 3D printed interfaces were developed for older children and adults. The physiological sensors (HR and EDA) explored during the co-design workshops performed well and would be beneficial for monitoring affect, however the large size of the EDA sensor was problematic. A conductive PLA used to 3D print the interfaces was tested as the possibility of printing a conductive surface would enable the surface of the interface itself to function as a skin conductance sensor. However, after exploring multiple conductive filaments, none were sufficiently conductive to function as an EDA sensor. It is possible to use other conductive materials to measure skin conductance rather than conductive filament, although when using conductive fabric as an EDA sensor the resistance was noticeably lower than the original EDA sensor. Due to the limited nature of conductive materials to function as a reliable EDA sensor, the original sensor in addition to the HR sensor were soldered to the microcontroller, creating the electrical components for the 3D printed interfaces as shown in Figure 3.5.

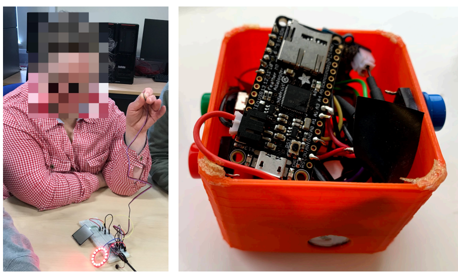


Figure 3.5: HR sensor connected to a breadboard used during the co-design workshops (left) compared with soldered sensors within the 3D printed cube (right).

During the prototyping stage a ball was first developed, this provided ample space to include all of the electronics but also presented its own set of challenges. The foam filling of the ball resulted in any haptic sensations being lost meaning only visual feedback was included using multi-coloured LEDs. A Soft teddy was explored next which presented similar challenges to the ball, different designs were explored including adapting the prototype demonstrated to participants although the design of this soft toy made it challenging to measure touch. Instead, a new design was selected and the sensors were successfully embedded to measure touch and motion interactions. Finally, a soft cushion was developed that may provide comfort similar to the soft toys but for older children or adults. The conductive fabric based conductance sensor was implemented within the cushion to test its real-world usability. However, embedding all of the sensors within a cushion was more challenging than the toys requiring careful holes to be cut for the buttons, cables, LED to show when the cushion was on, and power switch. The ability for a cushion to measure skin conductance may simplify the process of collecting physiological sensor data but the sensor was less reliable when tested due to the conductive fabric losing connection with the microcontroller.

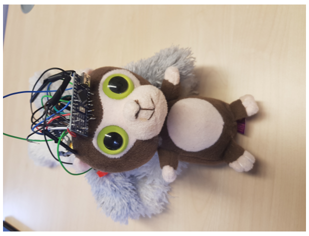


Figure 3.6: Early prototyping with soft toys.

A range of different shapes were considered for the 3d-printed interfaces as shown in Figure 3.7. After considering the various shapes, cubes were initially developed along with the torus selected during the co-design workshop. Initially the cube was printed with sharp edges and was too large to comfortably hold in one hand. The design was changed to include rounded edges and was made smaller while remaining possible to include all of the sensors. The size of the holes for the sensors and buttons had to be carefully designed to ensure they would be usable when encased in the interfaces. The placement of the physiological sensors within the interfaces was also given particular attention to ensure the user's thumb and fingers would be ideally placed to rest on the sensors. Learning from the development of the cube the other interfaces were simpler to develop, a range of interfaces were printed with some including less sensors to make the interface smaller such as a cuboid and others exploring new designs such as the pebble to make the interface easier to hold.

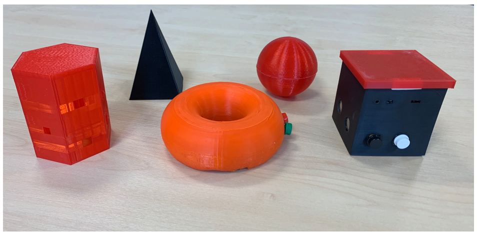


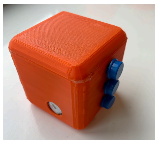
Figure 3.7: Shapes considered for 3D printed interfaces.

Fidgeting tools were also suggested for the interfaces. When experiencing poor mental wellbeing, people often fidget with objects as fidgeting is a natural response that helps regulate stress [211], [141]. Previous research shows squeezing interactions are preferred by children when angry but boredom was the most prevalent emotion to trigger fidgeting and clicking was preferred when bored [65]. This demonstrates that the fidgeting buttons enjoyed during the co-design workshops are a beneficial addition and should be embedded within affective TUIs resulting in the development of tangible fidgeting interfaces.

All of the sensors were connected to either an Atmega32u4 based microcontroller due to its small size or a raspberry pi 3 due to its more capable processor. A battery was also connected with each of the devices providing over 10 hours of continued use, in addition to an SD card reader to record all of the data.

Overall, nine prototypes were developed including both 3D printed shapes and soft interfaces. A soft ball and teddy were designed to cater for children, while a soft cushion embedded a large EDA sensor created from conductive material to enable its performance to be evaluated. The 3D printed interfaces included 2 cubes as these are easy to hold, a cuboid, a sphere containing sleeves to place fingers within, a spheroid designed to ensure the user's thumb will rest on the HR sensor and their palm will touch the EDA sensor in addition to the torus shape selected to print during the second co-design workshop. The sensors within each device are shown in Table 3.4.

Table 3.4: Description of the 9 developed wellbeing TUIs and the embedded electronics for real-world sensor data collection.

Device	Image	Description
Ball		A soft ball embedding 9-DOF IMU to measure motion and capacitive sensor to measure touch with multi-coloured LEDs to perform visual feedback
Cube (touch)		A 3D printed cube embedding 9-DOF IMU, capacitive touch, HR and EDA sensors with haptic feedback
Cube (buttons)		A 3D printed cube embedding 9-DOF IMU, fidgeting buttons, HR and EDA sensors with haptic feedback
Teddy		A soft teddy embedding a 9-DOF IMU and capacitive touch sensor with visual feedback
Torus		A 3D printed torus embedding HR, EDA, 9-DOF IMU and capacitive touch sensors with haptic feedback
Cushion		A soft cushion embedding EDA, 9-DOF IMU and capacitive touch sensors with haptic and visual feedback
Cuboid		A 3D printed cuboid embedding 9-DOF IMU and capacitive touch sensors with visual feedback
Sphere		A 3D printed sphere embedding HR, EDA and 9-DOF IMU sensors with haptic feedback
Pebble		A 3D printed spheroid embedding HR, EDA and 9-DOF IMU sensors with haptic feedback

3.5 Discussion

This co-design study has explored the design and development of TUIs to monitor real-world affect. A co-design methodology was adopted based on previous research with participants whose mental wellbeing can often be diagnostically overshadowed and who commonly have difficulty in expressing their emotions [102]. This co-design process promotes the role that people with intellectual disabilities have in acting as collaboration partners.

This research has many implications for both affective recognition technologies and the process of co-designing technologies with people who have intellectual disabilities. The five developed affordances aimed to increase individuals' autonomy [187] and promote communication within the workshops. It is hypothesised that participants' investment in the co-design process resulted from their ability to recognise the practical applications of affective technologies whilst also appreciating the impact that they may have on the daily lives of diverse user populations.

When co-designing with people who have intellectual disabilities it is vital to gradually unfold their creative potential to encourage meaningful participation [199]. Therefore, a number of activities were mediated with education professionals and conducted within each workshop that were designed to elicit design input and opinions, reach consensus and check understanding, such as the prioritisation of requirements.

A number of probes (instructional and conversational instruments) were also used during the co-design process to aid inclusivity [291]. Design probes were used such as demonstrations and hands on experiences with existing prototypes, sensors and actuators to serve as conversational instruments. During conversations design probes were iterated in conjunction with co-design participants, for example it was explained to participants that including all sensors in each design would lead to a TUI that was physically large. Furthermore, all information such as notes and video recording was collected in a structured way for subsequent analysis and major outcomes were recorded on a note board in picture form and simple sentences as a joint record of achievement, that was easy to understand for the entire design team.

Mixed probes were also used as the co-design session lasted from morning to early afternoon, allowing probes to incorporate physical (drawing, prioritising cards) and digital (3D modelling, exploration of electronics) elements and provide continuity between exploration and prototype testing. This meant that ideas captured from the whole team could be storyboarded and 3D printed all in one day, to make concrete the connection between design decisions and embodiment to support cognitive accessibility.

During the ideation phase, the production of prototypes was completed in a stimulating and playful environment, free from pressures that intimidate and block the creativity of the participants, including the use of design kits (drawing materials, cards, electronic components and sensors) as facilitating tools. The participants have previously been involved in the design of enabling technologies and hence they bring experience and expectations to be central to design decisions at co-design workshops.

These probes along with the five developed affordances and the inclusion of analysis and validation phases ensured participants were able to fully participate and provide valuable feedback. The feedback gathered from the co-design process resulted in the development of multiple affective interfaces for real-world monitoring.

3.5.1 Research Questions

The following sections use the analysis from Section 3.4.1 and feedback from the evaluation session where participants experienced the final developed interfaces to answer the four research questions.

3.5.1.1 How can the co-design process be adapted to best suit the needs of those with intellectual disabilities?

Co-designing for people with severe, profound and complex intellectual disabilities has shown additional challenges as not all tasks were successfully completed by all participants - such as drawing a new interface design. Significant communication challenges were also present. However, the majority of the tasks were completed successfully, particularly the interactive activities within the co-design workshops, involving exploring sensors and feedback mechanisms and the 3D printing of new interfaces to immediately show participants the results of their design decisions. These interactive activities helped maintain engagement, improved understanding of new concepts and improved communication within the workshops. By combining interactive sessions with storyboarding to gain feedback, it has been possible to gain valuable insights, aiding the design and development of future affective interfaces.

Overall, the co-design approach adopted addressed the limitations experienced by people with intellectual disabilities (e.g. communication and working memory), enabling them to participate more effectively. This approach takes a practical stance in guiding how co-design methods can be made to work in realistic settings and adjusted to the needs of participants who experience intellectual disabilities.

3.5.1.2 What are the optimal design guidelines for prototyping affective tangible interfaces?

The results from this study show there is no one-size-fits-all solution to the design of mental wellbeing interfaces. Instead, different devices should be developed for different age ranges with soft devices for young children and 3D printed interfaces for older children and adults. However, all devices should remain inconspicuous and not appear as medical devices to help reduce stigma.

During the evaluation participants stated that the interfaces developed thus far are suitable for potential users. During the time participants experienced the devices they stated that their use made them feel happy, in particular the fidgeting buttons helped them feel calmer. Teachers and end users agreed, liking the shapes of the 3D printed interfaces in addition to the ubiquitous nature of the devices with the sensors being embedded within objects such as cushions and teddies. Overall, the devices were found to be suitable for their intended purpose of monitoring affective state and collecting real-world data.

3.5.1.3 Which sensors would be most beneficial for user interaction?

A range of non-invasive sensors were explored using co-design sessions to measure physiological changes and physical interactions. Physiological sensors measuring HR, HRV and EDA present the greatest opportunity to automatically monitor wellbeing due to their correlation with the sympathetic nervous system and non-invasive nature. However, the analysis shows touch and motion sensors should be included within all devices due to their simplicity and ease of use, with the more complex physiological sensors reserved for 3D printed interfaces for older children and adults. Overall, the sensors included within the developed interfaces were considered suitable to collect real-world sensor data and affective state.

3.5.1.4 How do users believe on-device feedback could help improve their mental wellbeing?

A range of feedback actuators to serve as interventions and improve wellbeing were explored including haptic, visual and auditory feedback in addition to fidgeting tools. The co-design workshops helped establish the requirements such as the preference for fidgeting buttons and haptic feedback to improve wellbeing. Participants believed these feedback mechanisms have great potential to improve wellbeing with some participants finding the fidgeting aspects of the final tangible fidgeting interfaces especially relaxing. This demonstrates the use of the device and the fidgeting buttons themselves can serve as a benefit of using the artifact without the need for additional feedback mechanisms.

Visual feedback patterns were also enjoyed by users but it is not possible to use this feedback to convey information to users with intellectual disabilities and whilst the use auditory feedback was suggested to play stories and calming sounds, the quality of the sound is of high priority. Visual feedback, slow haptic feedback and fidgeting buttons were found most relaxing and have been embedded within the developed prototypes to act as real-time interventions.

3.6 Conclusion

Inclusive co-design workshops and focus groups have been conducted to rethink the user design approach of affective TUIs. Adjustments to traditional co-design techniques included demonstrations, real-time 3D printing, prioritising cards and interactive electronics to enable successful and practical co-design with people with intellectual disabilities. In particular, the 3D printing of new interfaces improved engagement, ensured participants understood the discussed technologies and demonstrated how their decisions influenced the final designs. Thematic analysis of the qualitative data outlined many recommendations and resulted in a range of new interfaces being developed. In the future additional methods of analysis could continue to be explored such as Interpersonal Process Recall [43] where the recorded co-design session is played back to participants as a stimulus for recall and reflection, allowing them to expand upon or clarify their comments. Also, the use of fictional inquiry [291] could be explored to imagine how co-designers with intellectual disabilities may use their devices to express and communicate their emotions and how this process might support their future design decisions. Overall, the participatory process has enabled the successful design and development of TUIs for real-world data collection and affect recognition.

Chapter 4

Tangible Techniques for Real-World Labelling and Collection of Wellbeing Sensor Data

To build an accurate and reliable affect recognition system a large labelled dataset is first required to train the model. However, collecting a real-world labelled dataset is a challenging proposition as sensor data must be labeled at the point of collection. Even though real-world physiological sensor data is notoriously challenging to collect due to body movements impacting sensor data [52] and the challenging propositions of labelling at the point of collection, it is vital to collect data in different settings, and not just controlled 'synthetic' data from experiments to ensure the model and device can successfully operate in the "wild" (naturalistic settings). This chapter seeks to address the challenges of labelling real-time sensor data by exploring on-device methods of labelling for inclusion within the developed tangible interfaces, resulting in the collection of a real-world labelled affective dataset. This section is adapted from [352], previously published in Personal and Ubiquitous Computing.

4.1 Introduction

Labelling data is not a trivial task, especially as the promise of affective TUIs is to make the possibility of ubiquitous wellbeing inference a reality. However, the labelling of sensor data is essential to enable classification models to be developed. Real-time emotions are often reported using the Ecological Momentary Assessment (EMA) [284], however this requires answering multiple questions often using a mobile app which is not suitable for labelling sensor data while using a TUI. Currently, real-time sensor data labelling is an unwieldy process with limited tools available and poor performance characteristics that can lead to the performance of the machine learning models being compromised. So far, most of the attention has been focused on the processing power of edge computing devices [108] [223] and little attention has been paid on how to obtain clean and efficient labelled data to train models [178].

The techniques used to label data vastly vary dependant on the data type as images can be labelled using an automated process based on clickthrough data, greatly reducing the effort required to create a labelled dataset [315]. Crowdsourcing labels is often employed for images and audio data tagging as it is most commonly processed offline [322]. Web based apps have been developed that enable people from around the world to highlight and label objects within images [258]. Outsourcing the labelling of image, video and audio data is gaining popularity although this is not possible for time series data as activities cannot be deduced from the raw data meaning real-time labelling techniques must be developed [179].

When collecting data in the real-world, outside the confines of the research lab a participant could be doing anything from driving a car to eating in a restaurant. A hybrid data collection approach is most suitable when collecting sensor data due to the subjectivity of the data. Using a hybrid data collection approach allows self-reporting to be combined with the passive collection of sensor data [273]. Smartphone applications are a popular method to label real-time data although recently the use of new smartphone labelling techniques such as NFC and volume buttons have shown to be intuitive and popular when using an application is inconvenient [360]. It is crucial to label sensor data in real-time, because unlike images and audio it is not usually possible to label the data after the point of collection without the real-time context. Longitudinal data collection poses even greater challenges as it relies on multiple users continually self-reporting, while simultaneously wearing sensors for extended periods. These challenges may result in limited labelled data which significantly impacts model accuracy.

The labelling rate of sensor data can dictate which labelling approach to choose as data that frequently changes may require a higher labelling rate along with a more convenient approach. The sample size is another factor that can dictate labelling approach as the labelling of images can be automated or crowd-sourced whereas a large sample size of sensor data requires recruiting many participants for an extended period. The best approach to label data often fundamentally depends on the data and source type being recorded. Labelling at the point of collection is highly accurate as it is real-time, cost effective, time effective and enables in-situ sensor data to be collected. Thus far however labelling at the point of collection has had limited use mainly consisting of smartphone applications and its feasibility and performance has not been evaluated. There are numerous scenarios where labelling sensor data at the point of collection would result in the most effective and accurate data but there is currently no established framework.

The use of TUIs embedding different labelling methods present significant opportunities for the real-time labelling of sensor data. These interfaces can embed a variety of sensors enabling the collection of in situ data. TUIs can vary in size and shape but the developed interfaces from the co-design workshops contain ample space to include the necessary sensors in addition to a real-time labelling mechanism. On-device labelling simplifies the process of self-reporting by not requiring additional materials, such as questionnaires, and its ease of access promotes frequent labelling. When providing participants with tangible interfaces to collect a wide array of sensory data, embedding a labelling method directly into the device simplifies the labelling process. This concept creates a simple, tangible, easy to use method to label sensor data in real-time and in-situ aiming to improve the quantity and reliability of labelled data.

This chapter presents two main contributions. The first is the development of LabelSens, a new framework for labelling sensor data at the point of collection, promoting the adoption of labelling techniques that can achieve higher performance levels. Five prototypes are presented utilising different tangible labelling mechanisms along with a comparative mobile app. A pilot study aims to explore the impact different labelling techniques embedded within TUIs has on the accuracy of labelling, label rate, usability and classification performance. Using the data from the study a comprehensive performance comparison and analysis of the prototypes is then provided. Looking beyond the data collection stage, the classification accuracy of different labelling techniques is examined. The second contribution is the collection of a real-world affective dataset. Using the results from the experimental LabenSens study, tangible labelling techniques are embedded with the co-designed tangible interfaces to enable the real-time labelling of affective state. The methods used and challenges encountered when collecting real-world labelled physiological data are then explored.

4.2 LabelSens Feasibility Study

Labelling data at the point of collection provides many benefits including low cost, reduced time and the ability to label in real-world environments. TUIs present many opportunities to embed unique physical labelling techniques that may be easier to use than comparative virtual labelling techniques frequently used. Furthermore, TUIs provide ideal interfaces to directly embed a magnitude of sensors, negating the need for participants to carry the sensors in addition to a separate labelling mechanism.

TUIs can vary in shape and size ranging from small wearables to physical devices that are designed to be frequently interacted with. This enables a wide array of opportunities to embed sensors within a variety of objects which when combined with machine learning classifiers could be used to infer behaviour change, emotions, movement and more. However, before machine learning models can be trained labelled data is first required. By embedding a labelling technique along with the sensors within TUIs it ensures the sensor data and label are both being collected in real-time aiming to improve data collection rates and accuracy. This novel approach to in-situ labelling provides an easy to use interface that facilitates the collection of real-time labelled data.

Figure 4.1 demonstrates the concept of the LabelSens framework; pairing time series sensors with a physical labelling technique inside a TUI to collect in-situ labelled sensor data.

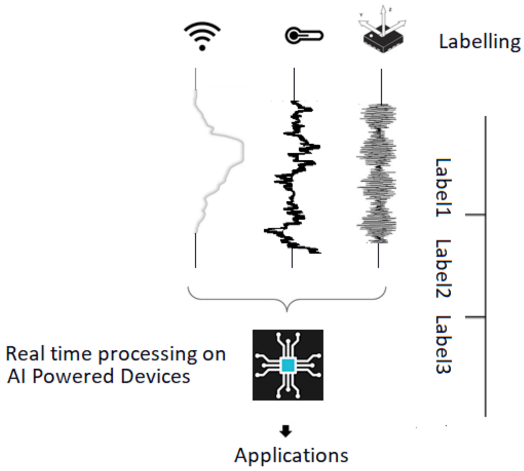


Figure 4.1: LabelSens framework: real-time sensor data fused with a label.

4.2.1 LabelSens Method

An experiment has been conducted that aims to explore the feasibility of self-labelling tangible techniques and the accuracy in which tangible techniques can label sensor data. The controlled collection of physiological data where people experience a variety of different emotions using each of the labelling techniques would be extremely challenging, inconsistent and time consuming. Instead, an initial pilot study has been conducted where tangible labelling mechanisms have been used to label physical activities from accelerometer sensor data which is more easily replicable whilst still requiring real-time labelling. This study enables each user to test all of the labelling mechanisms by completing different physical activities, not requiring users to experience different emotions to test the labelling approaches. Five new prototypes are presented that each contain a unique labelling technique along with a comparative mobile application that will be used to label human activities (walking, climbing downstairs and climbing upstairs) as these activities are frequently used in human activity recognition studies [252], [18], [9], [251], [160]. The developed labelling techniques are as follows:

- Two adjacent buttons (press one button for climbing upstairs, press the other button for climbing downstairs and press both buttons simultaneously to record walking)
- Two opposite buttons (press one button for climbing upstairs, press the other button for climbing downstairs and press both buttons simultaneously to record walking)
- Three buttons (one button each for climbing upstairs, climbing downstairs and walking)
- FSR to measure touch (Light touch for walking, medium touch for climbing downstairs, hard touch for climbing upstairs) with LED to visualise the label selection
- Slide potentiometer (slide to the left for climbing downstairs, slide to the middle for walking and slide to the right for climbing upstairs)
- An Android mobile application provided on a Google Pixel 3 smartphone with 3 virtual buttons to label walking, climbing downstairs and climbing upstairs.

Each TUI is a 6cm * 6cm * 6cm 3d printed cube that contains a labelling technique combined with the required sensor and microcontroller. The size of the TUI could be reduced dependent on the labelling technique used and the sensors required but here all interfaces were the same size to reduce bias. The embedded electronics include:

- Arduino Nano microcontroller. Due to its small size, open source nature and wide compatibility with sensors.
- IMU to record motion data. An IMU with 9 degrees of freedom has been used as it integrates an accelerometer, a magnetometer and a gyroscope to provide better accuracy and additional data.

- Micro SD card reader to locally record the IMU sensor data along with the user inputted label.

It is envisioned that TUIs will be used to label a maximum of 5 classes to ensure users are not overwhelmed and can sufficiently label at all times. Additional buttons could be added e.g. 1 button each for up to 5 classes but as only 3 activities are being classified the impact of having varying number of buttons (2 or 3) can be explored. The mobile app provides visual feedback once a label has been selected and presents easier opportunities to include more labels but users may still be overwhelmed by numerous virtual buttons. The buttons and slide potentiometer also enable users to easily visualise the activity they are labelling and feel tangible feedback whereas when using the touch sensor it is difficult to distinguish between the three levels of force. To visualise the selected label a multicoloured LED has also been incorporated into the touch device that changes from green to yellow to red when the device is touched with low, medium and high force. Figure 4.2 shows the electronic circuit and the developed TUI for the three buttons and slider interfaces.

The mobile application was developed for the Android operating system and was tested using a Google Pixel 3. The application consisted of three virtual buttons in the centre of the screen labelled "*downstairs*", "*walking*" and "*upstairs*". When a button is pressed the text at the top of the screen changes to show the currently selected label. Finally, at the bottom of the screen are two additional virtual buttons to begin and end the recording of data. The sensor data along with its label is then saved to a CSV file stored on the phone's internal storage. A challenge when developing the mobile app for data labelling is the frequency of the data as the gyroscopic data had a significantly lower frequency than the accelerometer data resulting in the reduction of data sampling frequency.

This pilot study involved ten participants using all of the developed interfaces containing the 5 different labelling techniques and mobile app while undertaking *walking, climbing upstairs* and *climbing downstairs* over a set route to ensure sufficient data was collected for all three activities. Participants were instructed that the label should only be recorded when commencing a new activity and if an incorrect label is recorded then the correct label should be recorded as soon as possible to simulate real-world labelling. Ideally the labeling system should

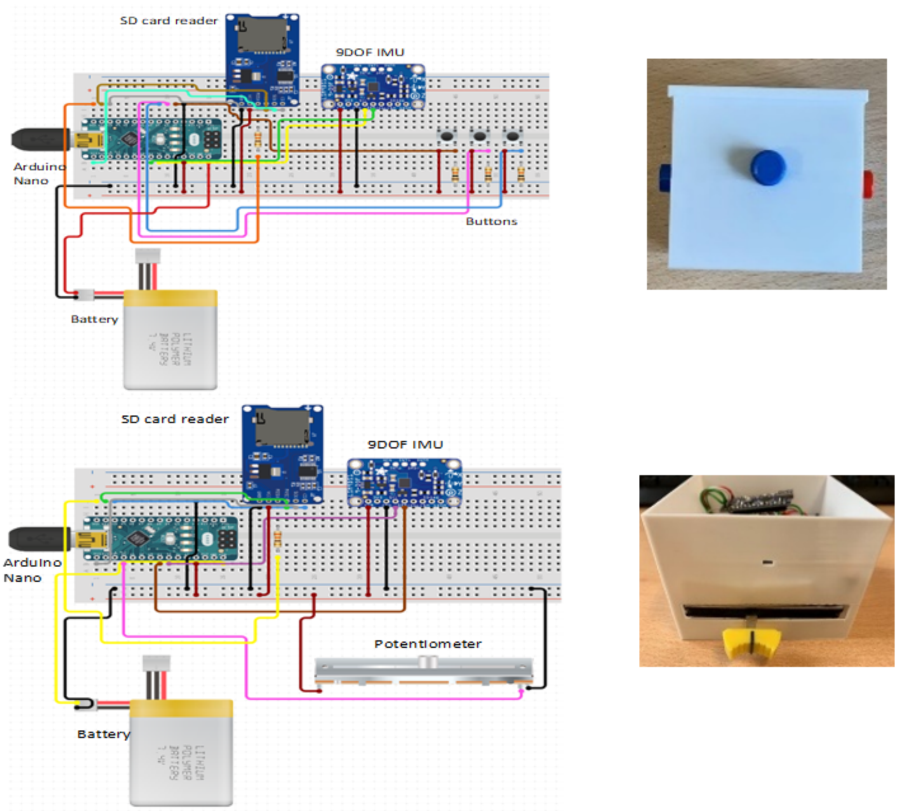


Figure 4.2: Example of two electronic circuits and interfaces with three buttons and slider labelling mechanisms.

be unobtrusive; in a way that the process of labeling the data should not alter or affect the data being collected. Therefore participants were not accompanied during the data collection period to realistically simulate in-situ data collection which is the fundamental purpose of these interfaces. No issues arose during data collection with each participant understanding how to use each of the interfaces and successfully collecting data from all devices. The three activities allowed for each participant to experience the different labelling techniques as well as collect sensor data which can be used to examine the accuracy and performance of each labelling technique.

4.2.2 Results

4.2.2.1 Labelling Rate

The maximum labelling rate of the devices is a key factor in deciding a labelling technique as some forms of sensor data can frequently change, requiring a new label to be recorded multiple times every minute. To measure the maximum rate at which it is possible to label data a preliminary experiment was conducted where participants were instructed to use each interface continuously for 2 minutes to record the maximum number of label changes possible. Participants were instructed to label each of the available labels as much as possible to assess the ease of use of labelling using each interface. Figure 4.3 shows the total number of times each label was recorded on each of the devices during the preliminary experiment.

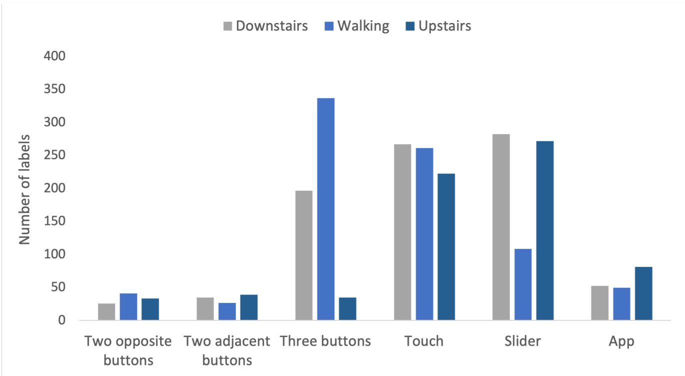


Figure 4.3: Maximum labelling rate for each label per device during the preliminary labelling rate experiment.

The devices with only 2 buttons showed the lowest data rate for each of the three labels because of a delay that was required to prevent mislabelling when simultaneously clicking both buttons to record the third label. The delay ensures that if a user releases one button slightly before the other when pressing both buttons to record the third label, the third label will still be recorded rather than the label for the button released last. The app shows a higher labelling rate than the devices with two buttons but is not significantly greater due to the difficulty in pressing virtual buttons that can easily be missed in comparison with physical buttons.

Three buttons resulted in significantly more data recorded although very little data was recorded for one of the buttons possibly due to the third button being more difficult to reach as each button is located on a different face of the cube. The touch sensor recorded a high labelling rate for all three labels as to reach label 2 (high setting) by pressing the sensor the user must first record label 0 and 1 as they increase the force exerted. The slider demonstrated high labelling rates for label 0 and label 2 but not label 1 because it is easiest to slide the slider from one end to the other but the slider was rarely located in the middle of the device long enough for the label to be recorded. This shows the touch and slider techniques are easy to label the extreme values but intermediary values are more challenging to frequently label. If all labels need to be frequently labelled then buttons may be the best labelling technique although the position of the buttons can greatly impact the ease at which labelling can occur.

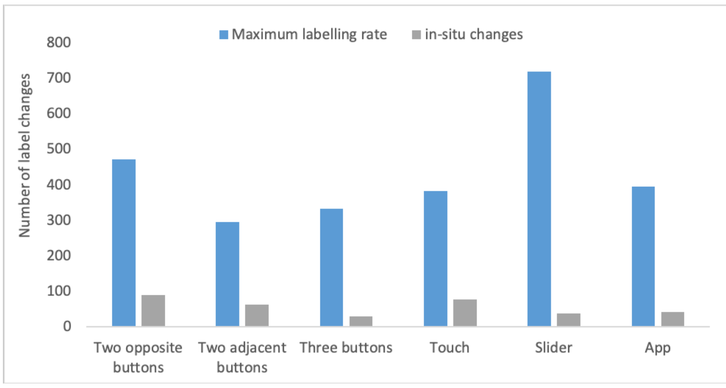


Figure 4.4: Comparison of total maximum label changes vs in-situ label changes per device.

It is also vital to compare the number of times the label changed over the 2-minute period to evaluate how simple it is to change label for each technique. Figure 4.4 shows the slider recorded the most label changes overall because of the simplicity to navigate between the labels, followed by two opposite buttons which is surprising due to its low labelling rate. This demonstrates that while the use of buttons does not result in the highest labelling rate it is simple to switch between the different labels and should be used when the label will change frequently. Touch, three buttons, the mobile app and two adjacent buttons all performed similarly well showing there is little difference in accessing all of the labels when using these devices.

Once all participants used each device to label *walking, climbing downstairs* and *climbing upstairs* the data was extracted, enabling comparisons to be established. The route participants were instructed to follow would require 11 label changes if followed correctly. Figure 4.5 shows the average number of label changes compared with the correct number of label changes. The average rate at which labels were changed from one label to another during the collection of physical activity data shows three buttons recorded fewest in-situ labelling changes at 5.6 significantly lower than the correct number of label changes (11) which is surprising but potentially shows the difficulty in accessing all of the buttons. Two opposite buttons had the highest overall rate of in-situ labelling changes with an average 17.6 labels recorded. Labelling via touch had a consistently high rate of label changes for users but this again could be due to the requirement of looping through all of the labels to reach the desired label. The mobile app achieved a slightly higher rate than three buttons and slider but still lower than the actual number of labels that should have been recorded showing the potential benefits of tangible labelling methods. Overall the slider and three buttons produced the lowest rate of label changes during data collection showing these labelling techniques should not be utilised with data that requires frequent labelling changes because of the difficulty in accessing all three labels. Two adjacent buttons recorded the most accurate number of labels showing the ease of accessing the buttons increased the accuracy of labelling.

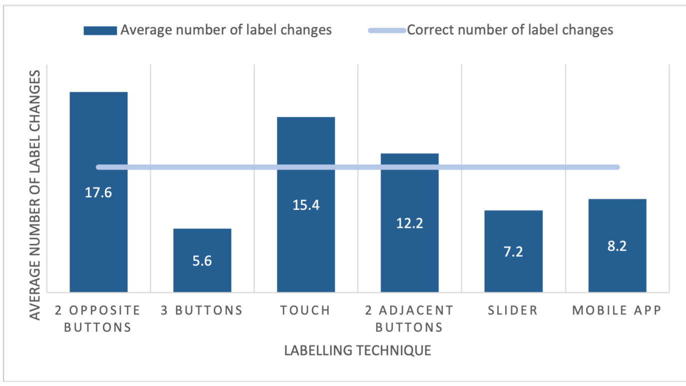


Figure 4.5: Total number of in-situ label changes per device.

Table 4.1 shows the total number of in-situ recorded samples from all participants for each of the devices. Touch and slider have the highest total number of samples recorded as when using these labelling techniques each label must be cycled through to change the label. Two opposite buttons has the smallest number of labels which is to be expected as a delay had to be added after a button press to prevent incorrect labelling. Because of the delay it was expected that two adjacent buttons would similarly have a low data rate but it achieved a higher rate than three buttons, possibly, because of the difficulty in accessing the three different buttons on different faces of the cube. This shows the position of the buttons has a greater impact on the number of labels recorded than the number of labelling interfaces embedded into the device. The comparative mobile app performed better than the buttoned devices but not as well as the slider or touch interfaces demonstrating the benefit of TUIs when a high labelling rate is required. While all interfaces recorded more walking labels than any other label as expected due to the route having more walking than stairs, the app had the fewest downstairs labels recorded demonstrating the difficulty in accessing virtual buttons. Similarly, two adjacent buttons had a smaller proportion of upstairs and downstairs labels which is surprising as these labels are the easiest to access (by clicking a single button) compared with labelling walking that required both buttons to be pressed simultaneously. It is also likely that the touch and slider interfaces have more downstairs samples than upstairs samples as downstairs must first be cycled through to reach either the walking or upstairs label. Overall, the dataset collected is large enough to train classification models and a sufficient number of samples have been collected from all three classes.

4.2.2.2 Deep Learning Classification

In order to identify the three activities from the sensor data collected, deep neural networks have been used to develop three predictive models. The performance of the three supervised, deep learning algorithms were tested to classify the sensor data into the three activity classes. A multilayer RNN [235] with LSTM [130], a multilayer RNN with GRU [142] and multilayer RNN with a stacked LSTM-GRU were selected due to their high performance and capabilities in classifying time

Table 4.1: Number of in-situ samples collected per labelling interface for each of the 3 labels.

	Downstairs	Walking	Upstairs	total
Slider	5910	10828	3657	20395
Two Adjacent Buttons	2441	9364	2132	13937
Touch	3188	14551	2442	20181
Three Buttons	2197	6688	2635	11520
Two Opposite Buttons	1537	2214	2285	6036
App	2066	10324	4316	16706
Total	17339 (19.5%)	53969 (60.8%)	17467(19.7)	88775

series data.

The dataset collected from each of the five interfaces and mobile app was used to train the three models over 10 epochs with 10-fold cross-validation. The 9 degree of freedom sensor data used as input data for the model was collected using a sample rate of 7Hz. The initial learning rate of the model was set to 0.0025, using a batch-size of 32 and an overlapping window size of 100 with an overlap of 20. Figure 4.6 shows the accuracy of each model. The stacked LSTM-GRU displayed little impact compared with the LSTM. Meanwhile, the GRU outperformed the LSTM and stacked models for most labelling techniques with the exception of two adjacent buttons where the LSTM network achieved the highest accuracy of all the labelling techniques at 92.8%. The overall GRU accuracies ranged between 68.5% and 89% demonstrating the impact different labelling techniques have on a dataset and thus the accuracy of a classification model.

The two adjacent buttons labelling technique achieved the highest accuracy of all the devices which is unexpected due to its complex nature where 2 buttons represent 3 labels. The second most accurate device, touch, was also unexpected due to the more complex interaction required of pressing the device using varying levels of force to record the different labels. It is possible that the more complex action forced users to have a greater focus on labelling their activity resulting in more accurate labelling. This however may not be sustained if the device was to be used for several days. Even though three buttons and the slider labelling techniques resulted in the lowest changing labelling rate, they achieve consistently

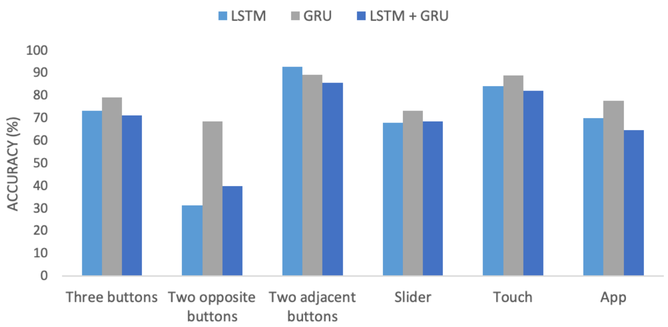


Figure 4.6: Comparison of deep learning techniques on the combined data collected from each devices.

high accuracies in the three trained models. This demonstrates that although it may be more difficult to collect fast changing data with these techniques, the collected data is reliable and capable of producing accurate classification models. The mobile app again performed moderately achieving 77.8% accuracy which although is not as high as touch, two adjacent buttons or three buttons it is greater than the slider and two opposite buttons interfaces.

Figure 4.7, shows the accuracy and loss of the combined user test data for all of the labelling interfaces during each epoch when trained using the GRU model. The loss for each of the models gradually decreases but the loss for the touch and slider decrease significantly as would be expected due to these interfaces achieving the highest overall accuracy.

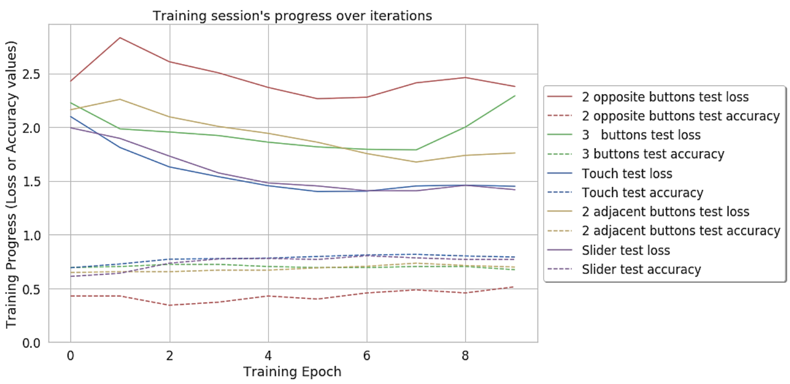


Figure 4.7: Comparison of training accuracy and loss when using GRU on the total data collected for each device.

It is possible that the datasets may contain potential biases for example if one user was particularly poor at labelling with one device it may significantly impact the quality of the training dataset. To evaluate potential bias and explore differences between users, the GRU model was used to train subject independent models from the data collected using each interface as shown in figure 4.8.

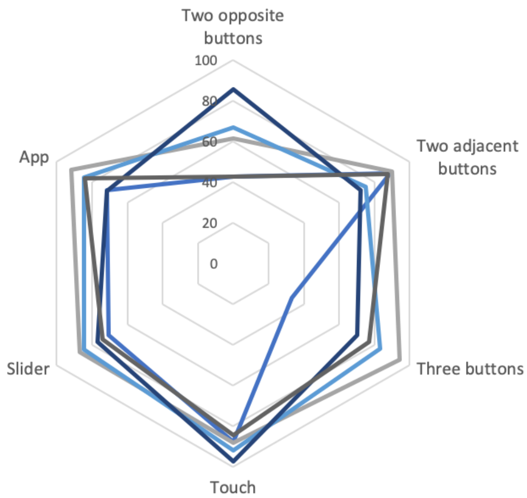


Figure 4.8: GRU model accuracy when individually trained on the first 5 users' data.

There are extremely wide variations in model accuracy ranging from 33.3% to 97.1%. Two opposite buttons and three buttons demonstrate the widest variation in model accuracy with accuracies reduced to 42.9% for user 1 using two opposite buttons and 33.3% for the same user using three buttons. As the lowest accuracies were all performed by the same user it indicates that this user experienced more difficulty using the interfaces than the other users. However, two opposite buttons also demonstrated poor accuracy (42.9%) when trialled by user 5, thus it shows that this interface results in poorer data collection as the models from the same user achieved consistently high accuracies for all other interfaces ranging from 74.2% to 87.5%. When comparing average model accuracy for each user it shows some users can result in significantly better data collection and therefore model accuracy, for example the overall accuracy between all interfaces for user 2 was 85.4%. The mobile app, two adjacent buttons, touch and slider all achieved high levels of accuracy when tested with each user's data demonstrating the reliability for those interfaces to consistently collect accurately labelled data. The touch interface achieved the highest overall accuracy at 97.1% when trained using data collected by user 4, although the data from the other interfaces collected by user 4 did not result in as high accuracy demonstrating that user preference and familiarity with an interface plays an important role in the quality of data collected.

Classification accuracy alone does not provide an informed overview of the most beneficial labelling technique. The F1-score, a harmonic average of the precision and recall, for each label and device has been calculated, as shown in Table 4.2. Overall, the walking label, has consistently higher precision and recall compared with the upstairs label which has the lowest F1-scores. The mobile app demonstrates good precision and recall when classifying *upstairs* but extremely poor precision and recall when classifying *downstairs*, potentially due to more mislabeling occurring when labelling *climbing downstairs*. The slider, two adjacent buttons and touch show the highest F1-scores which demonstrate their consistency as useful labelling techniques. Even though three buttons had a higher accuracy than slider, its F1-score is extremely low when labelling *up-stairs*, demonstrating its unreliability in classifying this class.

Table 4.2: F1-Score for each physical activity label when trained using each of the labelling interfaces.

	Downstairs	Walking	Upstairs
Slider	70%	82%	69%
Two Adjacent Buttons	82%	91%	75%
Touch	69%	94%	83%
Three Buttons	59%	80%	30%
Two Opposite Buttons	58%	75%	42%
App	23%	60%	82%

Cochran's Q test was performed to evaluate the three different models (L=3) for each labelling technique, providing a chi squared value and Bonferroni adjusted p-value. Cochran's Q test is used to test the hypothesis that there is no difference between the classification accuracies across multiple classifiers distributed as chi squared with L-1 degrees of freedom. Cochran's Q test is similar to one-way repeated measures ANOVA and Friedman's test but for dichotomous data as the classification will either be correct or incorrect and can be applied across more than two groups unlike McNemar's test [75].

Table 4.3: Cochran's test and F test comparing the developed classification models for each labelling interface.

	COCHRAN'S Q CHI^2	COCHRAN'S Q P-VALUE	F TEST	F-TEST P-VALUE
Slider	1.4	0.498	0.699	0.498
Two Adjacent Buttons	7.167	0.028	3.76	0.026
Touch	7.457	0.025	3.729	0.025
Three Buttons	6.143	0.046	3.136	0.046
Two Opposite Buttons	2.533	0.282	1.277	0.285
App	13.241	0.001	6.852	0.001

Assuming a significance level of α=0.05, Cochran's Q test shows for touch, two adjacent button, three buttons and the mobile app the null hypothesis can be rejected as all three classifiers perform equally well. For the remaining labelling techniques, the null hypothesis has failed to be rejected showing there is a significant difference for the classifiers on those datasets. The F test was also performed to compare the three classifiers as it is regarded analogous to Cochran's Q test. Assuming the same level of significance the slider rejects the null hypothesis in addition to two adjacent buttons confirming Cochran's results.

Cochran's Q test shows there is a significant difference between the three models when trained on the two opposite buttons and slider datasets but does not show where the differences lie. To see which models contain the significant differences the McNemar test was performed to compare the predictive accuracy of each model using the two datasets.

Table 4.4 shows the resulting p values when McNemar's test was performed. There is a significant difference between all of the models for both two opposite buttons and slider with the largest difference being between LSTM and the stacked network for both datasets. This demonstrates that both the labelling technique and the network architecture result in significant differences in the

Table 4.4: McNemar's test comparing the 2 opposite buttons and slider classification models.

	Two opposite buttons			Slider		
	GRU	LSTM	Stacked	GRU	LSTM	Stacked
GRU	NA	0.228	0.125	NA	0.286	0.596
LSTM	0.228	NA	0.546	0.286	NA	0.845
Stacked	0.125	0.546	NA	0.596	0.845	NA

models' accuracy and reliability.

4.2.3 Evaluation of LabelSens

To ensure the effectiveness of the labelling techniques it is also vital to gain users' preference. 50 users were asked which labelling technique they preferred. Figure 4.9 shows the results from the 50 users with 22% preferring three buttons as it was simple to understand and use due to there being one label per button although this labelling technique did not result in accurate models. Similarly 22% of people preferred two adjacent buttons with the mobile app following, which is surprising as majority of people are familiar with mobile apps so it would be expected to be the most popular. The users found three buttons and two adjacent buttons to be simpler to operate than the mobile app due to the physical buttons which are quicker and easier to press than the virtual button on the app which were often missed. Two opposite buttons followed again possibly due to the simplicity and familiarity of buttons to label data. The slider was well received but the granular control made the middle label more difficult to access meaning careful consideration had to be made to ensure actions were being correctly labelled. Finally, the fewest number of people preferred the touch based labelling technique due to the complexity of having to touch with varying levels of pressure to correctly label the data. However, touch did result in highly accurate models showing that while the increased attention required is not preferred it does ensure accurate data labelling but this may not be sustained over long periods.

While the user preference of labelling technique does not correlate with the accuracy achieved for each method it shows the benefits of using buttons as they are well-received by users and also achieved high classification accuracy. A lower

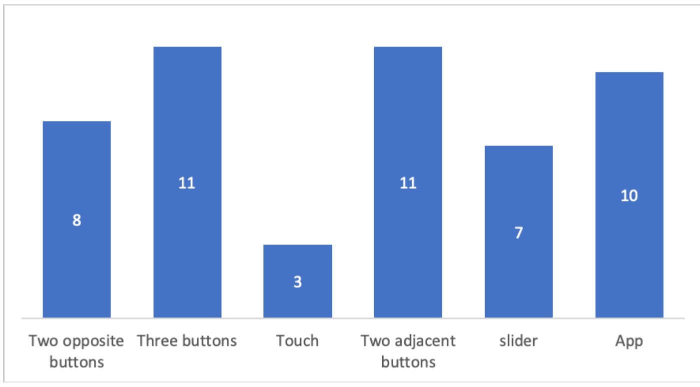


Figure 4.9: Comparison of 50 users' labelling preference.

number of buttons than labels is well received by users and achieves the highest accuracy, but the number of buttons must remain similar to the number of labels to ensure users do not experience confusion when labelling. The position of the buttons has also shown to impact user preference. In terms of labelling rate and model accuracy, two adjacent buttons were preferred by users and resulted in 24.3% higher model accuracy than two opposite buttons which had a higher total number of recorded in-situ labels but a lower labelling rate. It is imperative to balance user preference with the rate at which the data needs to be labelled and the accuracy required from the model when selecting an appropriate labelling technique.

Novel labelling methods including the slider and touch displayed their own strengths and weaknesses. Labelling using touch resulted in high model accuracy and labelling rate but was the least favoured by users. If accurate labelling is required for only short periods labelling via touch could be ideal. The slider was liked by users and had the highest labelling rate but achieved the second worse accuracy of all the devices at 73.4% showing the slider is best for continually changing or granular data that would be more difficult to label with buttons.

Surprisingly the mobile app was not the most popular labelling technique even though participants were more familiar with apps than the other interfaces. The data collected from the mobile app shows it achieved only a moderate labelling rate and model accuracy despite participants' familiarity. A possible reason why the mobile app did not result in the most accurate data is that virtual buttons can be easier to miss than physical mechanisms. However, when used in real world environments apps are easier to deploy but solely using an app does not allow for any additional sensors that are not embedded within the smartphone to be used. Apps possess many benefits when used to label motion data including ease of access but when additional sensors are required such as physiological sensors to monitor affective states using apps for labelling is not recommended over physical labelling techniques.

One of the most significant challenges encountered was the inconsistent quality of labelled data as when collecting in-situ data to train machine learning models it is not possible to ensure all users are successfully labelling their actions. By not accompanying users during the labelling process the experiment more replicated in-situ data labelling resulting in the different labelling rates experienced even though all users were instructed to walk the same route. Additionally, as users had to repeat the experiment five times to enable them to use each device, their labelling rate may change as they become more familiar with the experiment. To combat this, users were provided with the devices in varying orders preventing the same device from being used by all users at the same stage of the experiment.

4.3 Data Collection for Tangible Fidgeting Interfaces

The tangible labelling techniques developed pave the way for real-world labelled sensor data collection. The collection of labelled affective data is traditionally challenging as it data must be labelled at the point of collection. However, the developed affective tangible interfaces discussed in Chapter 3 and the evaluation of physical methods to label data demonstrate the capability to collect accurate real-world labelled sensor data.

A data collection approach has been co-designed with NICER group members and teaching support staff. The possibility of labelling real-world affective state from participants with intellectual disabilities creates many possibilities to reduce diagnostic overshadowing but also includes additional challenges regarding participants' ability to recognise and label their wellbeing. Previously, individuals with autism and Asperger's disorder have labelled their real-world emotions [161] and the development of tangible labelling techniques should further simplify the labelling process.

Initially, it was explained to all participants that before a computer model can infer their emotions, sensor data along with self-reported labels were required to train the models. Participant were provided with the opportunity to experience each of the tangible labeling techniques. All participants preferred buttons to the other labelling techniques stating *"I do think they would be better"* as they were the easiest to use to represent different states of wellbeing. Using different coloured buttons to label emotional states is easy to use for all participants but the number of labels would be limited to ensure simplicity for those with intellectual disabilities. While some participants suggested three labelling buttons for positive, neutral and negative affective states, the majority found this too complex and preferred the simple labelling option of two buttons representing positive and negative emotions as they are working at a developmental level where a simplification of emotions is required.

The findings from the exploration of tangible labelling techniques resulted in the inclusion of two buttons, one green and one red to label positive and negative states of mental wellbeing respectively, representing a single item wellbeing index

[309]. However, to resolve the limitations of on-device labelling mechanisms while ensuring the interfaces are still easy to use, the use of a paper diary was also used, allowing for a larger range of emotional states to be captured and to ensure the on-device labelling techniques are used correctly. Potential emotions and their representations for inclusion within the diaries were discussed and co-created, with emojis being favoured by the majority of participants to represent different emotions.

4.3.1 Experimental Setup

Six participants with intellectual disabilities each used one of the developed tangible interfaces incorporating the tangible labelling buttons over a two-week period in real-world environments. This differs from many previous studies that collected controlled experimental data or included specified activities during the data collection period to artificially impact wellbeing [194], [248], [290], [368], [289], [169]. Participants were instructed to use the red and green buttons to label negative and positive states of affect each time they used the interface. The label would then be stored along with the sensor data from each session of use. Participants were also instructed to correct their labelling as soon as possible if they accidentally mislabelled their wellbeing. In addition to being given a device to use in their home and work environments, each participant was provided with a diary to record their emotions as shown in Figure 4.10.

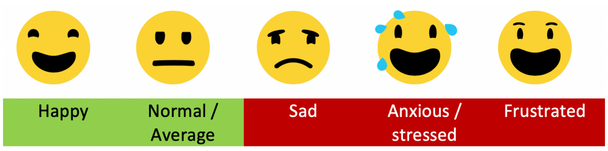


Figure 4.10: The emotions used in the diary enabling participants to record additional labels.

To ensure the diary was easy to use, five emotions (happy, neutral, sad, stressed and frustrated) were represented using emojis, ensuring emotions from each of the four quadrants from Russell's circumplex [259] were included. Each page of the diary consisted of three sets of emotions enabling labelling each morning, afternoon and evening of the data collection period.

4.3.2 Experimental Observations

After the two-week data collection period the devices along with the diaries were collected for analysis. One participant had to be immediately excluded due to the unexpected circumstance of their dog chewing the device. The remaining five participants completed their diaries and successfully used the device every day during the two-week period.

Participants used the interfaces for varying lengths of time with some participants using the device for a few minutes and others using it for several seconds at a time. Participants were instructed to label each session of use with some participants labelling before using the device and others after. Before the data could be explored it had to be cleaned to match the sensor data with the recorded label. When cleaning the data any sessions without a label were removed, if a label was recorded near a session (either before, during or directly after) the entire session was recorded with that label and if multiple labels were recorded in quick succession the last label was aligned with the sensor data as participants were instructed to correct the label if they accidentally mislabelled a session of use. If a label was recorded but no physiological sensor data was recorded (potentially due to fidgeting motions while interactive with the interface) then the accelerometer data was aligned with the recorded label but the physiological sensor data was removed. After cleaning the data all sessions of sensor data had a label aligned with them enabling further evaluation.

Upon evaluation of the diaries and the on-device labels it was shown that three of the five remaining participants were consistently reporting their emotion as happy or neutral throughout the two-week data collection period and never reported a negative state of wellbeing. One of the participants who recorded negative emotions only recorded two instances of such affect states during the two-week period, while the final participants added text to justify their selection of emotion and subsequently reported the most negative states. A potential explanation to the overwhelmingly positive recorded emotions suggested by an experienced teacher of students with intellectual disabilities is that participants with an intellectual disability are often keen to please researchers and give positive responses to questions. This demonstrates that the collection of overly positive labels is not due to participants not understanding the task but rather a pleasing desire which can be addressed through additional training.

When comparing the button labels with the diary entries some anomalies were discovered where the red button had been pressed to record a negative emotion, but a positive emotion was reported in the diary. This potentially shows the difficulty of using buttons to label data as participants may use the buttons as a fidgeting tool rather than their intended purpose to label the data. The participant who recorded two negative emotions frequently changed their label in a short period of time confirming that buttons were potentially purely viewed as methods of interaction rather than a labelling technique. At other times the same participant recorded a label for around 30 seconds then recorded a different label for over a minute showing either their emotion changed frequently, or it took an extended period for the participant to correctly label their wellbeing state. A potential challenge is that when users are stressed they may not pay much attention to correctly labelling their wellbeing states. Overall, the lack of negative emotions during this data collection trial in both the diaries and the on-device labels resulted in a biased dataset which is not sufficient to train machine learning classifiers and demonstrates that participants require further training on how to use the interfaces and labelling methods.

4.3.3 Evaluation of the Data Collection

After the data collection trial, members of the NICER group including those who participated in the data collection trial provided their feedback and evaluated the labelling methods. Participants stated they enjoyed using the devices with the majority taking it with them to use outside of their home. Some participants stated they placed it on their desk when at work or college so that it was always within reach and used it whenever they felt it was required as even without any technological interventions participants believed the act of fidgeting with the device helped improve their wellbeing. Other participants carried it with them in their pocket or bag showing the portability of the interfaces helps increase engagement and promotes device use.

The use of the two embedded buttons to label the data was considered appropriate and easy to use. However, nearly all participants found the stressed and angry emojis used in the diaries difficult to understand. The use photo realistic images of faces [26] was proposed, which may be easier to understand than emojis. Alternatively, it was suggested to change the colours of the emoji faces, where the happy emoji would be green and the angry emoji red.

The frequency of the self-reporting was also problematic as the diary only allowed participants to record their emotions over long periods of time, such as one label for the morning or afternoon. However, when stress or anger was experienced it was only for a short or transient period, but as the individual felt happy for majority of the time frame it resulted in positive emotions being recorded. To resolve this issue it was suggested that the diaries could ask for participants' affective states to be recorded during the past hour as this should increase the frequency of emotional variation. Users could also be instructed to use the device whenever they feel stressed or angry as these emotions don't usually last very long. This should help increase the frequency of data collection related to poor mental wellbeing states being collected, allowing for the successful training of machine learning classifiers.

Additionally, it was suggested that parents or carers could help provide insights into the mental wellbeing states of participants by completing their own questionnaires to expertly label participants' affective states, and justifying each emotion reported. However, the presence of a paid carer may pressure users into responding more positively and if parents and carers only complete questionnaires infrequently they will not capture the transient periods of time where users may experience negative emotions. Further training will be required to ensure that all participants with an intellectual disability understand each emotion as there is the need to first train them in understanding and recognising their different emotions before going onto record these as labels.

4.3.4 Revised Data Collection

The data collection approach was revised to obtain more accurate labels. The development of new labelling diaries continued to be discussed and designed with the participants and education specialists through focus groups over two design cycles, resulting in two new diaries being created both removing the previous time frames. The first used photo realistic images of a person expressing five emotions: happy, calm, sad, shocked and angry [26]. The second diary continued using emojis but rather than emojis expressing specific emotions it displayed five emojis that scaled from happy to sad with the colour gradually changing from green to red as shown in Figure 4.11 based on the Self-Assessment Manikin scale [99]. Before commencing a second data collection trial to collect additional real-world labelled sensory data, the two new diary styles were shown to participants. After exploring both designs the majority of participants preferred the photo realistic diary and hence this was used for labelling all future data. This was slightly surprising as those with autism may display and recognise facial affect differently [158] although all participants correctly understood each of the represented emotions.

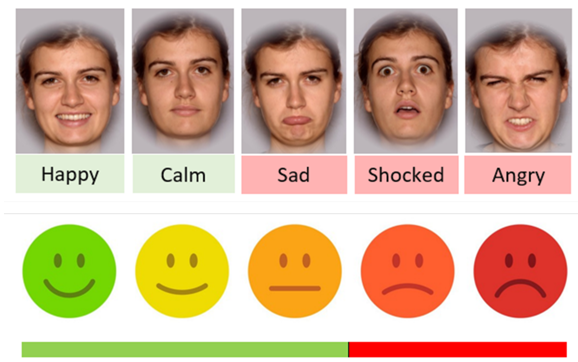


Figure 4.11: Updated data labelling diaries using photo realistic images (top) and five point emoji scale (bottom).

15 participants (9 Males and 6 females) were then asked to partake in a second data collection trial, 9 of whom have complex intellectual disabilities. Additional training on how to use the devices was performed with the help of education specialists including a head teacher with over 40 years experience in special education. Each of the emotional states were clearly explained to participants as there was previously difficulty establishing the difference between the stress and frustration states. These two states were subsequently changed to shocked and angry emotions which participants found easier to understand. Participants were also trained to correctly label their wellbeing by reaffirming their labelling was

private and that their honesty was valued, showing they didn't need to provide positive labels to please the researchers.

Participants were informed they were to use the device as frequently as possible throughout their normal daily life, in particular when experiencing a change in emotions, (labelling a minimum of 3 times per day), keeping the device with them for as much time as possible. Participants were also instructed to frequently label their affective states, especially when their emotion might change to ensure a wide range of labels connected to different affect states are collected as this was problematic with the previous data collection trial. Once the participants understood the requirement for increased labelling frequency, the interfaces were then selected by the participants themselves.

All participants used the interfaces during their daily life to collect real-world labelled affective data over a period of at least one week. After participants used the devices, the interfaces were collected so that the data could be analysed and used for computational analysis. The on-device labels have been used to explore the data as participants used the on-device labelling buttons more often than the diary due to their convenience. Also, when using the diaries participants regularly used only two emotions, one to represent negative emotions and one to represent positive emotions, matching the labelling buttons. Out of the nine participants with intellectual disabilities, the data collected from five participants was insufficient to train classification models as either the devices were not used sufficiently or all of the recorded labels were positive emotions resulting in biased datasets. The remaining participants all successfully used the devices to collect physiological data in addition to motion data and labels.

The number of positive and negative samples and percentage of positive and negative labelled events recorded from each of the remaining ten interfaces is shown in Table 4.5. The data shows users 1, 2, 3, 4, 5 and 8 used their devices the most collecting a large number of both positive and negative samples. Users 6, 9 and 10 collected mostly positive samples during the data collection period but also collected sufficient negative samples to train machine learning models. User 7 collected the least the number of samples, using the device the least although as both positive and negative samples were collected it remains possible to utilise this data.

Exploring the labelled events (each time a user recorded a new label) shows that all participants recorded a mixture of positive and negative labelled sessions with an average of 15 labelled sessions per user during the data collection period. Users 1, 2, 4, 8, 9 and 10 recorded more positive events than negative events with the highest ratio being 66.6% positive events and 33.3% negative events demonstrating sufficient negative events were still recorded. The duration of each session widely varied with participants who recorded fewer sessions (users 1, 2, 3, 7, 9 and 10) generally using the device for longer periods of time, often several minutes. Whereas participants who recorded more labelled events (users 4, 5, 6 and 8) generally used the device for shorter periods of time ranging from several seconds to a few minutes. This shows the interfaces were used in two distinct ways either frequent short usage sessions or less frequent longer sessions although both interaction methods resulted in a number of positive and negative events allowing for the data to be used to train classification models.

Table 4.5: Comparison of positive and negative real-world emotional state samples for all 10 users.

	Positive samples	Negative samples	Positive label events	Negative label events
User 1	30765	33482	66.6%	33.3%
User 2	29206	32760	66.6%	33.3%
User 3	18637	32002	50%	50%
User 4	80743	8400	60%	40%
User 5	27394	24253	50%	50%
User 6	4928	1355	43.7%	56.3%
User 7	211	796	50%	50%
User 8	12568	15184	52.9%	47.1%
User 9	9438	1329	60%	40%
User 10	15874	796	66.6%	33.3%

When exploring the real-world EDA data collected from the 10 users, Figure 4.12 shows the negative state of wellbeing has a clear wide distribution point whereas the distribution for the positive state of wellbeing is more dispersed, further demonstrated by the significantly lower median. Similarly, when exploring

HRV both states share the same distribution pattern but lower HRV has a much wider distribution when experiencing negative emotions and the median HRV is higher when experiencing positive emotions. This data demonstrates that the real-world physiological data collected from the tangible interfaces behaves as would be expected when experiencing poor wellbeing.

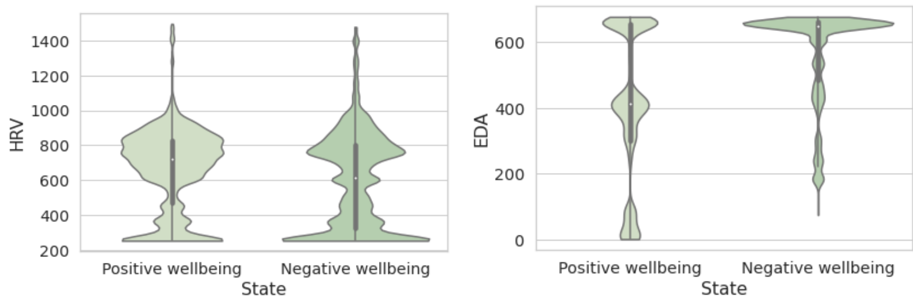


Figure 4.12: Comparison of real-world HRV data (left) and EDA data (right) collected using the tangible interfaces.

The additional training shows improved data collection results with labelled sensor data collected for most participants, although issues such as biased data and insufficient data from some participants were still encountered. When conducting wellbeing data collection trials with participants who have intellectual disabilities, it is vital to first ensure they have a full understanding of how to use the interfaces to record their wellbeing, as well as encouraging the frequent labelling of different wellbeing states to ensure the data collection is successful.

4.4 Conclusion

TUIs are ideal interfaces to infer affective state but first real-world labelled data must be collected to train the classification models. To address this issue and collect in-situ labelled sensor data five labelling techniques have been embedded into TUIs: two opposite buttons, two adjacent buttons, three buttons, slider and touch along with a comparative mobile application. The interfaces were used by participants to label three physical activities enabling the performance of each technique to be evaluated. It is vital to compare different labelling techniques as machine learning models can only be as accurate as the labelled data they are trained on.

During the pilot study participants used the six labelling techniques to collect data which was then used to train various RNNs. The results demonstrate that while a touch interface resulted in a high labelling rate and high model accuracy, it is the least favoured by users due to the high level of attention required to use the device. The mobile app was popular with users due to its familiarity but only achieved the fourth highest accuracy. The slider resulted in high user preference and labelling rate but poor model accuracy while two adjacent buttons achieved both high user preference and the highest model accuracy showing it is the most beneficial technique for sensor data collection. Based on these results 2 adjacent buttons was selected as the labelling technique to be embedded within the affective tangible interfaces as this approach simplifies the labelling of real-world sensor data.

Following the pilot study, the process of collecting real-world labelled affective data was explored. Two labelling buttons were embedded within the co-designed tangible interfaces due to their simplicity and effectiveness during the pilot study. The initial affective data collection trial demonstrated the challenges of collecting a real-world labelled dataset. After re-evaluating the data collection method, additional training was provided to users and by using on-device labelling, frequent accurate labelling was successfully completed during the second real-world data collection trial.

Overall, tangible labelling techniques within TUIs address many of the challenges facing the collection of in-situ, time series sensor data collection. Two adjacent buttons have been used within the tangible interfaces to enable users to label their mental wellbeing in real-time. This labelling approach encouraged the accurate collection of real-world affective labelled data that can then be used to train classification models.

Chapter 5

Mental Wellbeing Recognition using On-Device Transfer Learning

This chapter presents work applied to the real-world physiological wellbeing data collected using the developed tangible interfaces. This data is first leveraged to train 8 subject-independent deep learning models based on a leave-one-out cross validation approach. The use of TL is then explored to train individual affective models on-device, increasing model performance by developing personalised models using small labelled datasets. This section is adapted from [347], previously published in IEEE Sensors Journal and [350] previously published in IEEE International Smart Cities Conference 2021.

5.1 Introduction

The ability to infer affective states from sensors is an exciting proposition, as it could enable better real-world management of wellbeing. Non-invasive physiological sensors are ideal to infer affect as they can easily be embedded within TUIs, presenting a significant opportunity for wide role real-world adoption, unlike many previous approaches often using EEGs which are challenging to widely adopt outside of controlled experiments [8], [185]. The classification of mental wellbeing is a time series classification task which takes the physiological signals as input and outputs a label for each sequence. Deep learning presents many opportunities for advancing the classification of mental wellbeing as models can be trained using raw sensor data unlike machine learning classifiers that first require features to be extracted, which is often domain-driven and can be a time-consuming process. CNNs have traditionally been used to classify 2D data such as images but these networks are also employed towards extracting features from 1-dimensional sensor data as they can learn from raw time series data without first requiring feature extraction [107].

Affective states are often personal with individuals experiencing large variations in physiological parameters, despite experiencing the same state of mental wellbeing. Furthermore, those with intellectual disabilities often have their mental wellbeing challenges misattributed to their disability, thus making the realisation of a generalised model additionally challenging [102]. Advances in deep learning have helped increase affective modelling performance, however the majority of previous work has not considered the personalisation of models or the possibility of the models being used in different domains [98], [10], [340], [240]. For example a model developed using a controlled experiment dataset may not perform as accurately when used in real-world environments. Therefore, it is essential to develop personalised, subject-independent models for real-world monitoring as when working with a heterogeneous population there are numerous factors such as age, gender, disabilities and diet that result in a variation of physiological data [240].

Little research has explored the development of personalised affective models using non-invasive sensors that can be used in real-world environments. TL can help the development of personalised models by training a base model using labelled data from a different domain and transferring the learned knowledge to the new target domain [54], [233]. However, developing a custom model for each user remains a labour and time intensive task preventing real-world adoption. To address the challenge of real-world affective data collection and the development of personalised models an on-device TL approach has been devised. While the use of TL to improve modelling performance is not new conceptually, this TL approach for real-world physiological signal modelling helps address many of the traditional challenges by:

1. Exploring real-world, personalised affective modelling using few labelled samples from non-invasive physiological sensors. Thus, reducing the challenging proposition of longitudinal, real-world, labelled wellbeing data collection usually required to train affective models.

2. Removing the labour and time intensive process of developing personalised models by completing the TL process on-device using the custom-built tangible interfaces.

To achieve personalised affective models the development of an initial source model followed by the on-device TL approach to personalise the model is proposed. This approach hopes to alleviate many of the challenges traditionally associated with affective modelling such as the requirement of large datasets, enabling accurate real-world wellbeing inference.

5.2 Transfer Learning Framework

This work formulates model adaption as a cross-domain, personalisation TL problem. Here, a source model trained on a dataset collected from a controlled stressor experiment is adapted to classify real-world affective state using an individual's labelled physiological sensor data. While a TL approach may assist personalising models and improving performance across multiple domains, it has traditionally relied on collecting the target domain data in advance to then complete the retraining before the updated model can be exported. A new approach has been devised where the target user provides only a few labelled samples over a short period of time and TL is then performed to personalise the model on-device. Collecting a small labelled dataset and then personalising models on-device simplifies the process of developing personalised, cross-domain models as the device does not need to be returned for the TL model to be developed. It is anticipated that the personalised cross-domain models will improve the accuracy in which real-world wellbeing can be inferred. The following sections discuss the data collection for the source model and the real-world wellbeing data as well as the on-device transfer learning methodology.

5.2.1 Controlled Stressor Experiment (Source Model Data Collection)

Before the source stressor model could be developed, the stressed and relaxed physiological data first had to be collected. A lab-based stressor experiment has been conducted where participants' various emotions were stimulated using the Montreal stress test [74]. This experiment induced stress in 20 healthy participants aged 18-50 between June - September 2019 as approved by Nottingham Trent University human ethics board, application number 600. To allow the effects of stress and mental arithmetic to be investigated separately the experiment had three test conditions; rest, control and experimental. Each participant was initially briefed before completing a 3-minute rest period where participants looked at a static computer screen where no tasks were displayed. This was followed by 3 minutes of the control condition where a series of multiplication mental arithmetic questions were displayed which participants answered, followed by another 3-minute rest period. Participants then completed the stressor experiment where the difficulty of the questions increased and the time limit of the tasks was adjusted to be 10% less than the average time taken to answer questions during the training, taking it just beyond the individual's mental capacity. The time pressure along with a progress bar showing their progress compared with an artificially inflated average were both designed to induce stress during the 10-minute experiment. Finally, participants completed a 3-minute rest session and answered questions on their subjective experience of task load (NASA Task Load Index) [120], mental effort (Rating Scale Mental Effort) [371], emotional wellbeing (Self-Assessment Manikin) [38] and stress (visual analogue scale) [111].

Sensors: Participants wore hand-held non-invasive sensors on their fingers. The sensors recorded HR Beats Per Minute (BPM), raw HR amplitude, HRV, and EDA, each sampled at 30Hz to collect physiological data while experiencing relaxed and stressed states of mental wellbeing.

The subjective experience data from the post-experiment questionnaire is shown in Table 5.1. The results show that participants experienced high levels of mental and temporal demand, effort and mental effort but remained slightly positive with an average valence of 6.57. The average stress store was 5.36 which

Table 5.1: Controlled stressor experiment subject experience from the NASA Task Load Index (NASA-TLX), Rating Scale Mental Effort (RSME), Self-Assessment Manikin (SAM) and Visual Analogue Scale (VAS).

Type	Feature	Description	Average
TaskLoad (NASA-TLX)	Mental Demand (0: low - 10: high)	How much mental and perceptual activity was required (e.g. thinking, deciding, calculating, remembering, looking, searching, etc.)?	8.29
	Physical Demand (0: low - 10: high)	How much physical activity was required (e.g. pushing, pulling, turning, controlling, activating, etc.)?	3.07
	Temporal Demand (0: low - 10: high)	How much time pressure did you feel due to the rate or pace at which the task or task elements occurred?	6.69
	Effort (0: low - 10: high)	How hard did you have to work (mentally and physically) to accomplish your level of performance?	6.86
	Performance (0: poor - 10: good)	How successful do you think you were in accomplishing the goals of the task set by the experimenter (or yourself)?	4.21
	Frustration (0: low - 10: high)	How insecure, discouraged, irritated, stressed and annoyed versus secure, gratified, content, relaxed and complacent did you feel during the task?	4.93
Mental Effort (RSME)	MentalEffort (0: none - 10: extreme effort)	How high was the mental effort for the tasks you just finished?	7.07
Emotion (SAM)	Valence (1 - 9)	How do you feel at this moment? (unhappy - happy)	6.57
	Arousal (1 - 9)	How do you feel at this moment? (calm - excited)	4.64
Stress (VAS)	Stress (0: not - 10: very stressed)	How stressed do you feel?	5.36

is lower than expected especially when participants stated they experienced high levels of mental effort, although is higher than previous similar studies [169]. The low subjective stress scores could potentially be explained by participants misunderstanding the question and answering how stressed they felt at the time of completing the questionnaire rather than the experiment as multiple participants asked for clarification regarding this. Overall, participants experienced high levels of effort and mental and temporal demand during the experiment showing it is likely that high levels of stress were experienced.

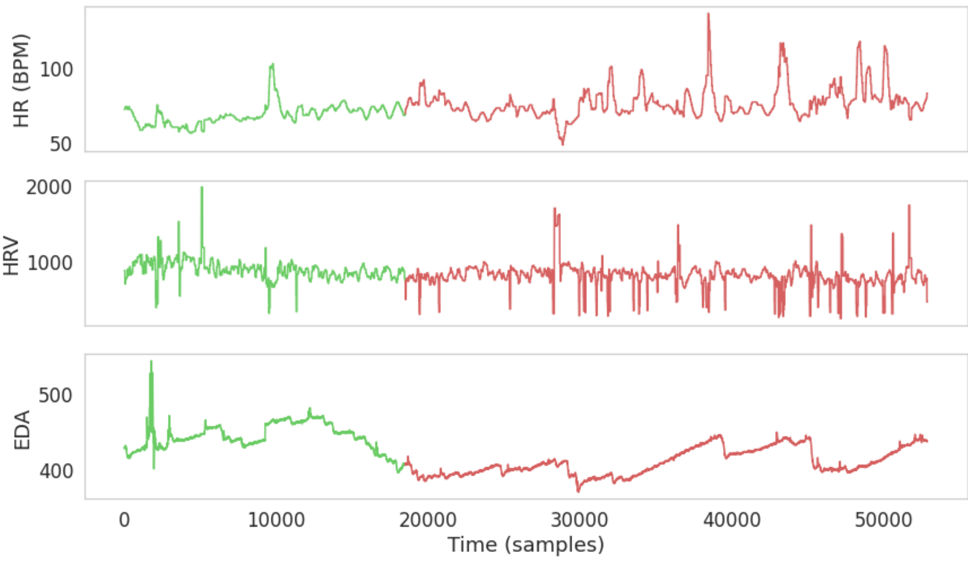


Figure 5.1: Comparison of one user's stressed (red) and relaxed (green) data collected from one the controlled stressor experiment for HR (top), HRV (middle) and EDA (bottom).

A similar number of samples were collected of both relaxed and stressed data, helping to reduce bias in the classification model. The controlled experiment dataset resulted in a total of 417251 sensor data samples when relaxed and 475232 data samples when stressed. The dataset contains HR (mean 79.4BPM, Standard Deviation (SD) 11.6BPM), HRV (mean 773.8ms, SD 152.6ms) and EDA (mean 320.4kΩ, SD 158kΩ) physiological sensor data. While time series data is traditionally challenging to classify by sight there are clear trends displayed such as when participants were resting the average HR was 76.9 (SD 10.5) compared with 81.6 when stressed (SD 12.1) showing an elevated HR when stressed. Similarly

HRV reduced from 795.4ms (SD 149ms) when resting to 754.9ms (SD 152.6ms) when stressed. Finally EDA significantly reduced when stressed from 345.3kΩ (SD 152.6kΩ) to 298.6kΩ (SD 159.5kΩ). Figure 5.1 shows a sample of stressed and relaxed physiological data for one user who completed the controlled stressor experiment demonstrating trends such as lower EDA and higher HR when stressed. However, this was not true of all participants with some experiencing more pronounced physiological changes than others, reinforcing the need for personalised models.

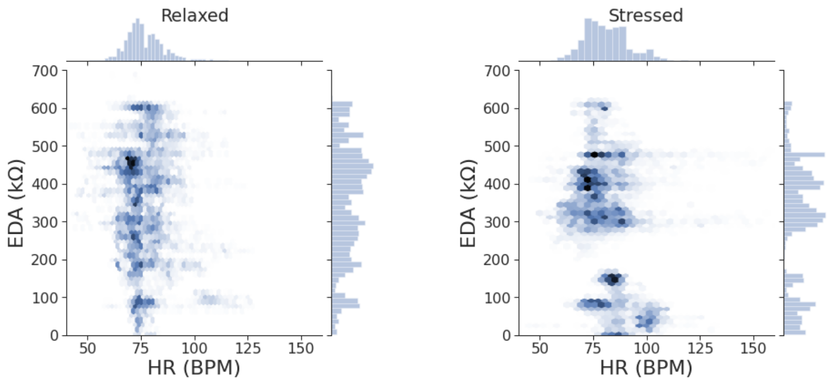


Figure 5.2: Summarising the controlled experiment EDA and HR data when participants were relaxed (left panel) vs data when participants were stressed (right panel).

To evaluate the physiological changes during the experiment, the HR (BPM) and EDA data is represented in Figure 5.2 as they demonstrated the greatest level of change between the rest and stressed states. Figure 5.2 shows that when stressed the distribution of the HR and EDA data is highly concentrated within two clustered areas compared with the relaxed data which is more sporadically dispersed. The dispersed data demonstrates that HR may still be high and EDA low when relaxed but this occurs much less frequently. Overall, the trends from the dataset show that EDA is an important indicator of stress and when paired with HR and HRV can show clear patterns of stressed and relaxed emotions.

5.2.2 Real-World Mental Wellbeing Trial (Target Model Data Collection)

The data collected from the data collection trial in Chapter 4, where a total of 15 participants were provided with a custom-built device for one week to collect real-world wellbeing data, has been used to enable the development of personalised models. The devices contained the same HR, HRV and EDA sensors sampled at 30Hz as used in the controlled experiment, in addition to accelerometers and gyroscopes to measure motion. These sensors were used due to their non-invasive nature, usability in real-world environments and direct correlation with the sympathetic nervous system. The on-device labelling method was used to label the sensor data. During the data clean-up process a nearby label (either immediately after or before) the sensor data was used as the label. If no label was recorded the data was removed and if a long session of use was recorded (potentially due to fidgeting while interacting with the interface) the data was manually analysed to see if multiple labels were recorded where the data could be split if not the entire session was labelled with the nearby label. If no physiological data was recorded but motion data was present the same process was completed to label the motion data and the physiological data was disregarded.

Nine of the fifteen participants had complex intellectual disabilities resulting in mental wellbeing issues often being diagnostically overshadowed by their disabilities, making a personalised solution extremely valuable. After examining and removing biased or insufficient data, the data from six participants with no disabilities (users 1, 2 3, 6, 7 and 8) and four participants with intellectual disabilities (users 4, 5, 9 and 10) has been used to examine deep learning affective models. The distribution of the data per participant is reported in Section 4.3.4.

5.2.3 On-device Model Personalisation

The stressed and relaxed data collected from 20 participants completing the controlled experiment has been used to train a source 1D CNN which will be adapted to perform the TL approach. The input data included HR, HRV and EDA data all sampled a 30Hz which was divided into segments of fixed lengths with an input vector of size 32 X 5. An overlapping sliding window strategy has been adopted to segment the time series data with a window size of 100 samples and a step of 20 chosen experimentally, by testing different window sizes from 10 to 400. The model was trained over 50 epochs with a batch size of 128, achieving 82.5% accuracy when classifying stressed and non-stresses states using hold-out validation with a 20% test split. The network architecture consists of multiple 1-dimensional convolutional layers followed by max pooling operations with a stride of 2 as shown in Figure 5.3. Batch normalisation layers are included along with a dropout layer with a rate of 0.5 to prevent overfitting before the softmax activation function. To compare performance, an LSTM network was trained using the same data for 50 epochs with a batch size of 128 achieving 65.5% accuracy, resulting in the 1D CNN being selected to perform the TL approach. This reduction in accuracy could be caused by LSTMs generally not being successful for short-time, frequently changing, and non-periodical data [167], such as data on wellbeing.

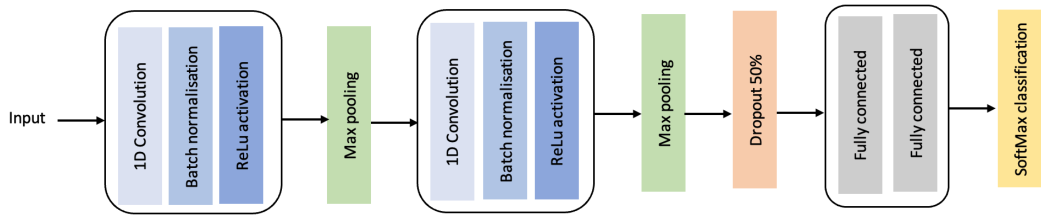


Figure 5.3: 1D CNN architecture.

The data collected from the tangible interfaces by the target users was used to personalise and adapt the source model to classify real-world positive and negative states of mental wellbeing as shown in Figure 5.4. The real-world data was labelled with the self-reported wellbeing states and when segments contained data from both classes a majority vote was used to label all of the data within the segment with the same label. The TL approach was applied on-device using each user's real-world data to re-train and adapt the source controlled stressor model. In order to achieve the transfer, the source network was frozen, the fully connected layer was removed and two fully connected layers were added, forming an adaption layer. The CNN was then re-trained on-device with the user's labelled data containing HR, HRV, EDA, accelerometer and gyroscope data sampled at 30Hz using the same window size (20) resulting in an embedded vector of size

160X9 before the fully connected layers.

The TL approach fine-tuned the models and updated the weights to develop personalised affective models. However, overfitting is a common problem when training CNNs using small datasets such as an individual's personal data and refers to a model that models the training data too well. Specifically, overfitting refers to a model that fits exactly against its training data therefore making it unable to accurately classify unseen data. This prevents the generalization of the model, limiting its use on new data. To help overcome overfitting, dropout layers have been included within the network that randomly ignore 50% of neurons during training [294]. Using dropout layers makes the training process noisy, forcing nodes within a layer to probabilistically take on more or less responsibility for the inputs therefore helping to overcome challenges when training CNNs using small personal datasets.

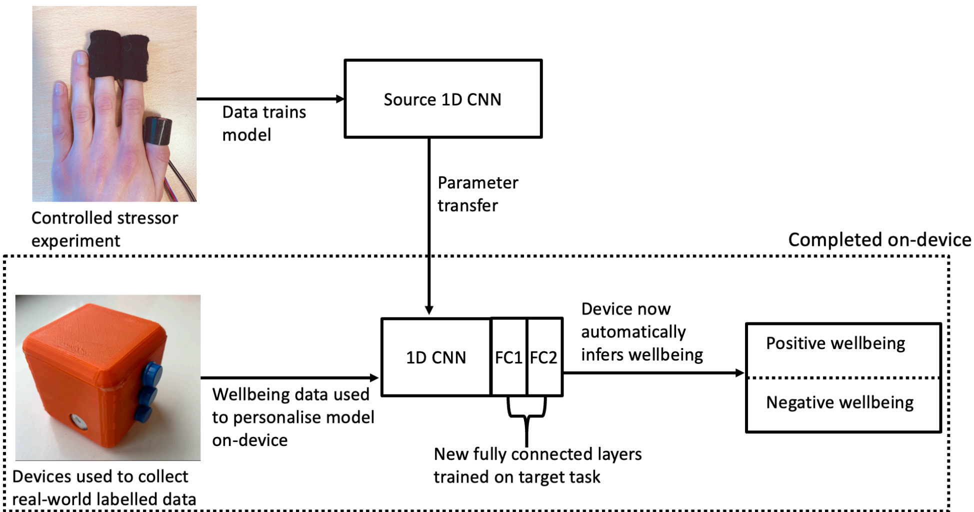


Figure 5.4: Transfer learning approach transferring learned knowledge from the source model trained using controlled stressor data to the target model classifying real-world mental wellbeing.

The TL personalisation approach where the model was re-trained with an individual's data to classify positive and negative states of affect was completed on-device using the custom built tangible interfaces. Each device contained a Raspberry Pi 3, with 1GB of RAM and a 1.2GHz quad core CPU for processing due to its affordability, compact nature and being sufficiently powerful to perform

TL. Using the Raspberry Pi it is possible to apply the TL approach, adapting the source model for the target user's real-world wellbeing data, on-device. This approach enables the model to infer real-world affective state rather than stress from the controlled experiment while simultaneously personalising the model for improved accuracy.

The on-device processing was slower than expected taking an average of 25 minutes to train the model using the TL approach due to the Raspberry Pi being resource-constrained. However, as this training only needs to be completed once, this is not a major limitation and remains quicker and simpler than returning the devices to train the model offline and then returning the device to users with the personalised model embedded.

5.3 Method

The real-world data consisting of HR (BPM), HRV, HR (amplitude), EDA and motion (accelerometer and gyroscopic) collected from the tangible interfaces was first cleaned and normalised before being used to train eight deep learning models to classify positive and negative states of wellbeing. Four of the tested networks (ResNet, TWIESN, Encoder and MCDCNN) have been previously explored using time series data [145] however, their effectiveness for affective modeling has not been explored. These four models have been compared with an additional four models (1D CNN, LSTM, CapsNet and Inception) to evaluate their affective modelling performance. Each model has been tested using Leave-One-participant-Out Cross Validation (LOOCV) to develop subject independent models, enabling comparisons with the personalised TL models. This method of testing accurately measures model performance on an individual basis and better simulates real-world performance.

In addition to comparing multiple deep learning classifiers the TL approach was also completed. Once each of the participants had used their device to collect a personal labelled dataset and the on-device TL approach had been performed, each user's individual model was tested using 10-fold cross validation. It is important to assess the performance of the model trained using a small dataset using a resampling technique such as k-fold cross validation as it allows for a better estimate of the performance of the model on unseen data, helping to establish whether overfitting has occurred.

5.4 Results

5.4.1 Multivariate Physiological Models

5.4.1.1 Deep Learning Models

The real-world physiological data (HR (BPM), HRV, HR (amplitude) and EDA) was used to train each of the aforementioned deep learning networks. The results for each model tested using LOOCV is shown in Table 5.2 where the highest accuracy for each user is highlighted.

Table 5.2: Comparison of 8 deep learning models' accuracy tested on individual users' physiological data.

User	1D CNN	LSTM	Caps-Net	Res-Net	TWIESN	Encoder	Inception	MCD-CNN
1	**74.1%**	64.2%	53.1%	67.8%	59.8%	47.9%	68.8%	47.9%
2	**85.1%**	68.4%	58.5%	59.6%	54.4%	47.1%	59.3%	56%
3	**72.5%**	66.7%	63.2%	35.9%	25.4%	36.8%	36%	36.8%
4	**85.5%**	56.4%	67.7%	33.5%	52.5%	9.4%	36.4%	9.5%
5	**86.2%**	74.7%	53.1%	51.1%	52.7%	53%	45.8%	53%
6	**83%**	55%	81%	24.8%	22.7%	21.6%	25.6%	78.4%
7	81.1%	50%	17.4%	**83.3%**	21%	21%	81.6%	21%
8	**85.3%**	68.3%	54.8%	40.2%	43.2%	45.3%	43.2%	54.7%
9	83.8%	59.9%	**88**%	47.4%	87.6%	87.7%	43.5%	87.7%
10	87%	71%	53.3%	27.1%	94.9%	57.9%	29.7%	**95.2**%
avg	**82.4%**	63.4%	59%	47.1%	51.4	42.77%	47%	54%

The results show that for the physiological model the 1D CNN outperformed all other models for the majority of users, achieving between 72.5% and 87% accuracy (SD 5.1) with an average of 82.4%. This shows that wellbeing is highly personal as there was a 14.5% variation in accuracy between users when tested using the same 1D CNN model. The 1D CNN achieved 19% higher average accuracy than the next best performing model (LSTM), which is surprising as RNNs are commonly used with time series data.

The results show a wide variation in accuracy between the models when using the same data. For the physiological models the average standard deviation between models was 20.1. Four users (4, 6, 7 and 10) experienced higher standard deviance between models than the average deviation, with user 7 experiencing the greatest deviation of 30.7 with a 65.9% variance in model accuracy.

Similarly, there was a wide variation between users when tested with the same model, with an average standard deviation of 17.96. CapsNet outperformed the 1D CNN by 4.2% for user 9, achieving an overall accuracy of 88%, while ResNet outperformed the 1D CNN for user 7 by 2.2% and MCDCNN achieved the highest overall accuracy of 95.2% for user 10. While there is a 14.5% range of accuracy for the 1D CNN, the two next best performing physiological models, LSTM and CapsNet show much wider variation of 24.7% for LSTM (SD 7.82) and 70.6% for CapsNet (SD 19.06). The model accuracy for CapsNet ranges between 17.4% to 88% and while the 17.4% accuracy is an outlier, all other models achieved higher performance with the same user's data. Similarly, model accuracy for MCDCNN ranged between 9.5% and the highest overall accuracy achieved, 95.2% demonstrating a wide range of 85.7% (SD 27.56). This demonstrates that while CapsNet and MCDCNN can be used to infer wellbeing, their high volatility results in inadequate subject-independent models.

5.4.1.2 Personalised TL Model

When performing the TL approach each personalised model was initially trained using only physiological sensor data (HR, HRV, EDA). However, it was not possible to apply the TL approach for user 7 due to this user recording the lowest number of samples which whilst sufficient to test using LOOCV was not sufficient to train the personalised model without extreme bias and overfitting.

The accuracies of the personalised models for the remaining nine target users ranged between 84% and the highest overall performing model at 99.6% with an average accuracy of 92.3%, as shown in Table 5.3. The Area Under the Curve (AUC) shows the degree to which the model is capable of distinguishing between the two classes (positive and negative states of wellbeing). The AUC scores range between 86.2% and 99.6% with an average of 91.9%, remaining similar to accuracy. The personalised TL model for each user outperformed their respective best performing non-TL model, demonstrating the potential for mental wellbeing to be more accurately classified using personalised models.

Table 5.3: Comparison of 9 target users' multivariate physiological model accuracy and AUC using the transfer learning approach.

	Accuracy	AUC
User 1	84%	86.2%
User 2	90.4%	90.6%
User 3	99.6%	99.6%
User 4	95.8%	91.6%
User 5	90.6%	90.7%
User 6	93.5%	92%
User 8	89%	89%
User 9	91%	95%
User 10	97%	92%
Average	92.3%	91.9%

To ensure the TL approach had successfully learned the new real-world domain, the original source model trained using the controlled stressor data was tested using the target users' real-world data. The average accuracy achieved was 48.77%, representing a significant reduction in accuracy. These results demonstrate the model developed to classify stress is not capable of classifying real-world affective state, as while stress is an example of a negative affective state there are numerous other negative affective states that elicit different physiological responses. These results confirm the TL approach has greatly improved cross-domain performance, enabling more accurate inference of real-world mental wellbeing.

The aim of using TL was to additionally personalise the models. To demonstrate model personalisation, each of the target users' models were used to classify the data from all of the other users. The results show an average 36.09% reduction in accuracy when a user's personalised model was tested with alternative users' data. This demonstrates the TL approach has developed cross-domain personalised models as there is a significant accuracy improvement when the personalised model is tested using the matching user's data.

5.4.2 Mutivariate Motion and Physiological models

5.4.2.1 Deep Learning Models

Motion data (accelerometer and gyroscopic) was also collected from the interfaces, however the motion data from three of the devices was corrupted and therefore not usable to train models, resulting in 7 users. Table 5.4 shows the results for each of the deep learning models when trained using physiological and motion data. The 1D CNN model again outperformed all other models for the majority of users, achieving an average accuracy of 71.7%, although there was an 10.7% reduction in average accuracy compared with the physiological 1D CNN model. However, the CapsNet model achieved the highest accuracy of both models for user 7 and TWIESN for user 9.

Table 5.4: Comparison of 8 deep learning models' accuracy tested on individual users' physiological and motion data

User	1D CNN	LSTM	Caps-Net	Res-Net	TWIESN	Encoder	Inception	MCD-CNN
4	**87%**	52.6%	9.4%	79.7%	54.5%	62.2%	28.9%	49.7%
5	**79%**	57.8%	53.1%	47.4%	53%	52.8%	50.3%	53.9%
6	**81.1%**	72.3%	81%	45.6%	21.7%	24.2%	41.7%	78.4%
7	72%	29.2%	**82.6%**	40.2%	25.2%	21%	79.2%	79%
8	**54.8%**	49.2%	45.2%	36.7%	45.2%	44.9%	39.9%	36.9%
9	50%	59.9%	12%	42.3%	**87.6%**	80%	42.8%	80.4%
10	78.3%	71%	50%	59.1%	**95.1%**	77.7%	49.3%	95%
avg	**71.7%**	56%	47.6%	50.1%	54.6%	51.8%	47.4%	67.6%

Combining the physiological data with motion resulted in overall reduced performance although increased accuracy for user 4 by 1.5%. When comparing the best performing model for each user between the physiological and combined physiological and motion models, it shows there is a wide variation between a 30.8% decrease and a 1.5% increase in performance with an average decrease of 5.66% for the combined models.

The combined motion and physiological models demonstrated high deviations between models, similar to the physiological models, with the average standard deviation between models being 19.6 ranging from 6.2 to 25.4. The high deviation between models when tested on the same data shows the importance model selection has on performance.

5.4.2.2 Personalised TL Model

To explore whether motion data has an impact on the personalisation of affective state, the motion data was combined with the physiological data to train the target personalised models.

Table 5.5: Comparison of 6 target users' multivariate physiological and motion model accuracy and AUC using the transfer learning approach.

	Accuracy	AUC
User 4	96.3%	86.9%
User 5	89.2%	89.2%
User 6	90.3%	86.6%
User 8	91%	91%
User 9	96%	98%
User 10	98%	99%
Average	93.5%	91.8%

The results in Table 5.5 demonstrate that by combining the physiological and motion data accuracies between 89.2% to 98% can be achieved with an average accuracy of 93.47%. The combined physiological and motion TL models outperformed the physiological TL models for four out of the six users (user 4, 8, 9 10) and all models again outperformed the non-TL equivalent 1D CNN, increasing average accuracy by 22.37%. User 9 demonstrated the largest increase in accuracy, increasing from 50% to 96% when using the personalised model. This demonstrates that unlike in traditional deep learning models the inclusion of motion data with the TL models increased model accuracy, potentially demonstrating the manner in which users interacted with the interfaces is unique between users and can be used to indicate changes in wellbeing state.

5.4.3 Univariate Models

5.4.3.1 Deep Learning Models

As the 1D CNN outperformed all other models for the majority of users, this model was further explored to examine the impact training using each individual data source has on performance. The 1D CNN was again tested on a subject-independent basis using LOOCV for each of the 10 users with either the HR (BPM), HRV, EDA or motion data used to train the models, as shown in Table 5.6.

Table 5.6: Comparison of univariate 1D CNNs accuracy tested using LOOCV on 10 individual's HR (BPM), HRV, EDA or motion data.

User	HR	HRV	EDA	Motion
User 1	60%	66%	**75%**	N/A
User 2	65%	62%	**86%**	N/A
User 3	63%	56%	**86%**	N/A
User 4	50%	56%	72%	**86%**
User 5	78%	78%	77%	**80%**
User 6	74%	75%	**77%**	73%
User 7	69%	66%	**75%**	28%
User 8	74%	75%	**80%**	27%
User 9	68%	**71%**	70%	69%
User 10	61%	59%	**77%**	74%
Average	66.2%	66.4%	**77.5%**	62.4%

The results show that high model accuracy up to 86% can be achieved using only one data source with EDA being the most accurate univariate model for the majority of users, achieving an average accuracy of 77.5%. This demonstrates the importance of using EDA sensors when inferring wellbeing. However, HRV was the highest performing univariate model for user 9 and while the motion models achieved the lowest average accuracy of 62.4%, it was the most accurate univariate model for users 4 and 5, demonstrating the possibility of inferring wellbeing from motion data alone.

Surprisingly the EDA univariate models for users 1, 2 and 3 all outperformed the comparative multivariate physiological 1D CNNs by 0.9%, 0.9% and 13.5% respectively. However, the average overall accuracy for the univariate models does not outperform the average accuracy of 82.4% for the multivariate model, demonstrating multivariate models are most applicable for the majority of users. Overall, these results show that while multivariate physiological models provide the highest accuracy for the majority, univariate affective models can improve performance for individual users.

5.4.3.2 Personalised TL Model

Developing personalised mental wellbeing classification models using TL from only motion data has also been explored as the inclusion of motion data with the TL approach in Section 5.4.2.2 increased model performance. The motion data captured from users 4, 5, 6, 8, 9 and 10 was used to perform TL and develop personalised models. The univariate motion TL models outperformed each of the non-TL univariate CNNs for all users other than user 5. The TL models achieved an average accuracy of 88.05% ranging between 72.4% and 97% as shown in Table 5.7, compared with the average non-TL univariate motion model accuracy of 62.4%.

Table 5.7: Comparison of 6 target users' univariate motion models using the transfer learning approach.

	Accuracy	AUC
User 4	89.1%	72.3%
User 5	72.4%	72.6%
User 6	82.8%	75.4%
User 8	90%	90%
User 9	97%	89%
User 10	97%	89%
Average	88.1%	81.4%

Figure 5.5 shows the confusion matrices for user 5, the worst performing univariate and combined physiological and motion models. The confusion matrices confirm there were misclassification errors for both classes, although the majority of errors misclassified negative states of wellbeing as positive.

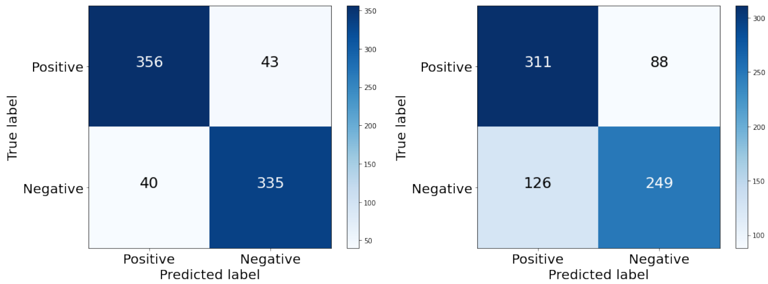


Figure 5.5: Comparison of confusion matrix for user 5 physiological and motion TL model (left) and univariate motion TL model (right).

While the motion univariate models were the worst performing TL models, surprisingly each user's model accuracy remained high when no physiological sensor data was used. These personalised models show the largest difference between accuracy and AUC, with the AUC on average 6.67% lower, demonstrating the greater level of misclassification errors. However, the motion univariate TL models outperformed the non-TL physiological univariate models in Table 5.6 for 4 of the users (users 4, 8, 9 10), further demonstrating the potential to infer wellbeing from motion data alone.

5.5 Discussion

The results demonstrate that overall 1D CNNs offer the highest affective modeling performance due to their low volatility and that a large dataset is not necessary to accurately infer real-world mental wellbeing. Instead, a small labelled dataset and TL can be applied on-device, addressing the challenging proposition of longitudinal affective data collection. The TL approach has shown to consistently improve model performance, demonstrating its ability to personalise affective models and increase cross-domain performance. Furthermore, its ability to infer the real-world mental wellbeing of those with intellectual disabilities may help reduce diagnostic overshadowing.

The highest accuracy was achieved using the TL 1D CNN fusing physiological and motion data, demonstrating the importance of using HR, HRV, EDA and motion data when inferring affective states. These TL models achieved an average accuracy of 93.47%, 22.37% higher than the comparative 1D CNN trained without the TL approach using LOOCV with significantly more samples, as shown in Figure 5.6. However, the accuracy of the models for users 5 and 6 decreased in comparison with the purely physiological TL models, showing that the inclusion of additional motion data does not necessarily increase model performance for all users.

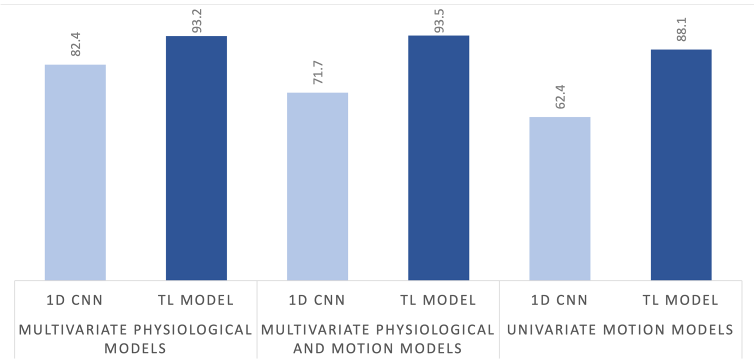


Figure 5.6: Comparison of average accuracy from 1D CNN models compared with personalised TL models.

Model accuracy remained similarly high when classifying mental wellbeing using only physiological data with the TL approach although it resulted in an average performance decrease of 1.57% in comparison with the physiological and motion TL models. The comparative non-TL CNN trained using the same physiological data resulted in a 9.5% average reduction in accuracy, again demonstrating the benefits of the TL approach. When the personalised models were trained with other users' data there was a 36.09% reduction in accuracy, confirming the TL approach successfully personalised the models. Furthermore, when the real-world datasets were tested using the source controlled experiment model there was a 43.55% reduction in accuracy confirming the TL approach also enabled the model to adapt cross-domain from controlled stress recognition to real-world affect recognition.

When using solely motion data to infer wellbeing, an average accuracy of 88.05% was achieved using the TL approach, outperforming all univariate non-TL models, although reducing average multivariate physiological and motion model performance by 5.42%. While this is lower than models trained using physiological data, it is higher than expected, with user 9 achieving their highest respective accuracy at 97% using the univariate motion TL model. Figure 5.7 shows the accuracy using the univariate TL model remained similar to the other TL models for users 8, 9 and 10 although there was a larger reduction in accuracy for users 4, 5 and 6.

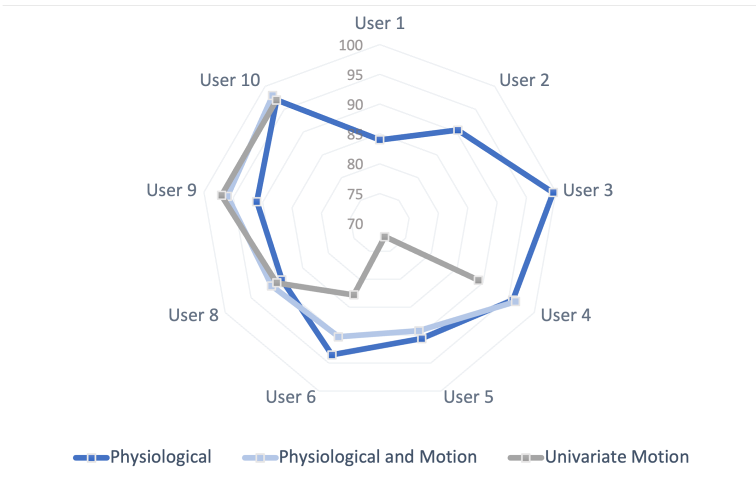


Figure 5.7: Comparison of accuracy for each user's TL model.

To investigate whether there is a distinct pattern of device motion between different states of wellbeing the total acceleration from the accelerometer data was calculated using 5.1, enabling exploration between the two states of wellbeing.

$$|\vec{a}| = \sqrt{a_x^2 + a_y^2 + a_z^2} \tag{5.1}$$

Figure 5.8 shows the distribution of the total acceleration across positive and negative states of wellbeing. The vast majority of total acceleration was between 1 and 1.2 where the negative state of wellbeing had a much wider distribution when the total acceleration was 1, had a smaller interquartile range and a lower median value. This demonstrates that the interactions with the interfaces changed with state of mental wellbeing, possibly as people fidget when experiencing poor wellbeing [211] [141]. This may be beneficial for future wellbeing devices such as wearables as they can infer wellbeing while remaining physically small by embedding fewer sensors.

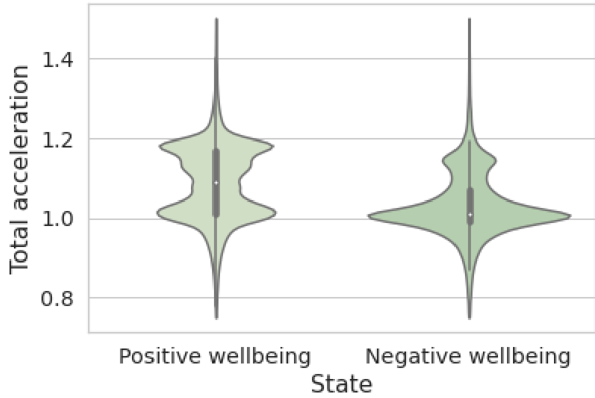


Figure 5.8: Violin chart comparing total acceleration between positive and negative states of wellbeing.

The ability to collect a labelled, real-world, personal dataset using custom-built devices and perform the TL approach on-device significantly simplified the process of developing personalised models. While the devices were mostly used as anticipated, the lack of negative wellbeing states reported by some participants resulted in their elimination from the study. However, as a personalised model was developed for each user, the requirement of a large dataset, which is traditionally challenging to collect was reduced.

The TL approach was successfully completed on-device using a Raspberry Pi and while the initial training of the personalised models took an average of 25 minutes, the models were then able to infer real-world mental wellbeing. This approach means future devices could be provided to the wider population to infer real-world wellbeing without the need to first collect large labelled datasets, greatly improving accessibility to personalised affective models. In the future, it would be beneficial to explore edge computing devices other than the Raspberry Pi to reduce the time required to train the model and perform the TL approach.

5.6 Conclusion

The personal, labelled, real-world data collected from the developed tangible interfaces has been used to train 8 deep learning subject-independent classifiers to infer affective state including CNN, CapsNet, ResNet, LSTM, TWIESN, Encoder, Inception and MCDCNN. Furthermore, a novel method for personalising affective models using a TL approach has been applied on-device to improve performance and adapt a source 1D CNN to infer real-world affect.

The results showed that the 1D CNN outperformed all other classification models for the majority of users, achieving an average accuracy of 82.4%. Univariate 1D CNN models, trained using a single data source were also explored, demonstrating EDA alone can achieve high performing models with an average accuracy of 77.5%. Surprisingly the univariate models for 3 users outperformed their comparative multivariate physiological models, demonstrating the potential to infer mental wellbeing from a single data source.

When applying the TL approach using physiological and motion data to classify positive and negative states of affect, an average accuracy of 93.5% was achieved. The TL approach not only enabled the model to adapt from a controlled stressor experiment to real-world affect recognition but also successfully personalised the models. Training a source model using a controlled dataset and then performing TL on-device, enabled high performing models without the challenging proposition of longitudinal real-world data collection.

Overall the TL approach only required a small sample of labelled data saving time, labour and money, while also improving model performance and operating across different domains. Advances in edge computing have enabled this personalised TL approach to be trained and run on-device, enabling the accurate inference of real-world affective state.

Chapter 6

Combining Deep Transfer Learning and Signal-image Encoding for Multimodal Mental Wellbeing Classification

It is impractical, time consuming and extremely challenging to collect large real-world datasets of individuals' wellbeing to train deep learning classifiers as it relies on users labelling their wellbeing for extended periods. To help address this issue a second TL approach has been developed to negate the requirement of having a large affective dataset to train deep learning classification models. This chapter explores signal-encoded images trained using a TL approach combined with an additional CNN trained using physiological sensor data, to improve affective modelling performance using limited data.

6.1 Introduction

Time series data from multiple modalities such as physiological and motion sensor data have proven to be integral for measuring mental wellbeing. However, the performance of machine learning models deteriorates considerably when training data is scarce, even more so with deep learning models. This lack of sufficient statistical power often hinders the progress of AI applications in monitoring and understanding wellbeing, since collecting longitudinal and annotated training data is very challenging [176]. This is due to the following reasons:

1. User availability, incentivisation and willingness to participate in longitudinal studies (or increasing study drop-outs beyond the first few months) [173]
2. Privacy, ethics and data protection issues [370], [239]
3. Data integrity and accuracy [195]
4. Cost and availability of monitoring devices [275]
5. Requirement to set up the device and extract the data by expert personnel needing specialized equipment [361]
6. Time consuming nature of real-time self labelling

In order to address these well reported problems, TL is often used [54]. Pre-trained models can be used to encompass methods that discover shared characteristics between prior tasks and a target task, reducing the necessity for large datasets [233]. Furthermore, previous research shows that fast changing, continuous sensor data such as accelerometer data can be transformed into RGB images which can then be used to train deep learning models [336]. Although the premise of presenting time series data as images is promising in extracting multi-level features and improving classification accuracy, most of the previous work only considered encoding univariate time series data as one image for a single channel of a CNN input [336], [357], [157]. This research proposes the combination of physiological sensor data with signal-encoded images in a TL model to tackle the challenging problem of monitoring the trajectory of wellbeing.

The use of signal-image encoding is explored to classify wellbeing using three techniques; GASF, GADF and MTF [336]. To address the limited sample size of wellbeing data a new CNN-TL-based approach has been developed to alleviate many challenges when classifying small datasets. This framework uses the signal-encoded images in a novel pre-trained TL model combined with a 1D CNN trained using physiological sensor data. Furthermore, the possibility of performing TL to classify stress from physiological data is explored by initially training a 1D CNN using a large physical activity dataset and then applying the learned knowledge to the target dataset. These approaches aim to use TL in addition to signal-encoded images to overcome problems with small training datasets, thus improving on the performance of conventional deep learning methods. Table 6.1 summarises the 2 proposed approaches.

Table 6.1: Description of the two proposed TL methodologies to overcome data scarcity.

1-Dimensional CNN TL Model	**Image Encoding TL Model**
1) Train a 1D CNN using a large time series dataset	1) Encode multivariate time series data as coloured images
2) Freeze the source model and remove the final layers	2) Leverage pre-trained object recognition models to apply a TL approach using the coloured images
3) Adapt the source model by training new fully connected layers with stressor dataset and updating the model weights	3) Train a 1D CNN to perform wellbeing classification from physiological data
4) Utilise the model to classify stress	4) Concatenate the pre-trained TL model with the 1D CNN to classify mental wellbeing

6.2 Data Collection

In this work, three different datasets have been used containing data from physiological, environmental and motion sensors, along with their labels. The first dataset uses physiological data from the controlled stressor experiment described in section 5.2.1, the second dataset shown in Section 6.2.1 collected physiological and environmental real-world data along with self-reported labels and the third dataset detailed in Section 6.2.2 is a publicly available human activity dataset containing accelerometer data [176].

6.2.1 EnvBodySens

Experiment setup: EnvBodySens is a dataset that has been previously collected by [202] that consists of 26 data files collected from 26 healthy female participants (average age of 28) walking around the city centre in Nottingham, UK on specific routes. The participants were asked to spend no more than 45 minutes walking in the city center. Data was collected in similar weather conditions (average 20°C), at around 11am. Participants were asked to continuously report how they felt based on a 5-point predefined emotion scale as they walked around the city centre experiencing general daily life stressors such as loud environmental noises and crowded environments. The 5-step SAM Scale for Valence from Banzhaf et al. [24] was adopted using a smartphone app developed for the study simplifying the continuous labelling process. The screen auto sleep mode on the mobile devices was disabled, so the screen was kept on during the data collection process. Data from six users were excluded due to logging problems. For example, one user was unable to collect data due to a battery problem with the mobile phone and another user switched the application off accidentally.

Sensors: The dataset is composed of non-invasive physiological data (HR, EDA, body temperature, acceleration) sampled at 8Hz, environmental data (noise levels, Ultra Violet (UV) and air pressure) also sampled at 8Hz, time stamps and self reports. The data was logged by the EnvBodySens mobile application on Android phones, connected wirelessly to a Microsoft wrist Band 2 [207] that was provided to participants to collect the physiological and environmental data.

The EnvBodySens dataset resulted in 29965 samples for state 1, 35333 samples for state 2, 106210 samples for state 3, 77103 samples for state 4 and 106478 samples for state 5. Figure 6.1 shows the EDA (mean 1455.2kΩ, SD 2870.5kΩ) and HR (mean 74.5BPM, SD 11.8BPM) for all participants when experiencing each of the five self-reported states of valence from 1 being most positive to 5 being most negative.

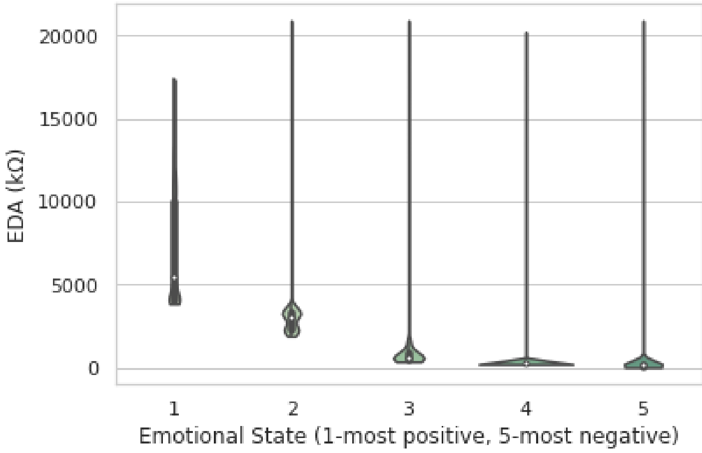


Figure 6.1: EnvBodySens EDA data for reported emotional states from 1 (positive) to 5 (negative).

The distribution in Figure 6.1 demonstrates that as users record poorer states of wellbeing, the average EDA value decreased. The EDA data collected behaves as expected with the median EDA value gradually decreasing as users experience worsening wellbeing.

Figure 6.2: EnvBodySens HR data for reported emotional states from 1 (positive) to 5 (negative).

However, Figure 6.2 shows wellbeing levels do not impact the distribution of HR like EDA; instead the distribution of HR remains relatively similar for all wellbeing states. Reported wellbeing state 2 has the highest distribution of HR reaching over 120 Beats Per Minute (BPM) even though this is the second most relaxed state. As users experienced worsening wellbeing the upper adjacent values are reduced, which is unexpected as when users experienced poor wellbeing they are more likely to have increased HR [304]. The outlier HR data in states 1 and 2 that go beyond 180BPM are most likely artifacts of the data due to sensor error, demonstrating that there is little change in HR over the 5 states of wellbeing. This demonstrates that HR alone as used in most commercial wearables, may not be sufficient to monitor affective state, requiring additional data modalities such as HRV and EDA. Overall, the EDA data behaves as expected while there is little to distinguish HR during the different states of wellbeing.

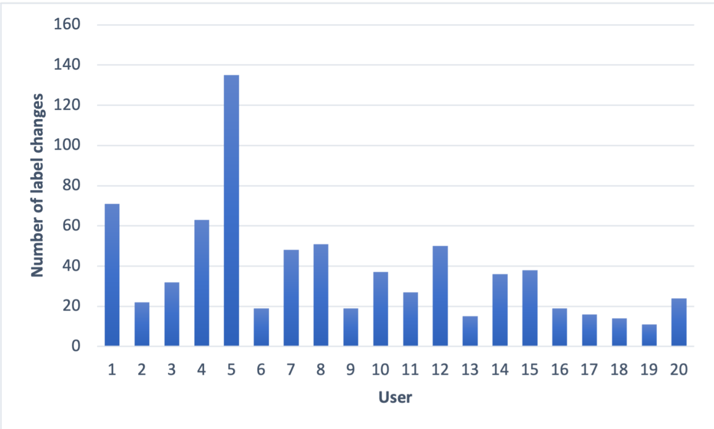


Figure 6.3: Total number of label changes per participant.

During the data collection process, 5345 self-report responses rated from 1 (most positive) to 5 (most negative), where sufficient samples for each rating were collected (1-8.44%, 2-9.95%, 3-29.91%, 4-21.71%, 5-29.99%). Data was successfully collected from all classes but class imbalance from an individual's dataset may impact the performance of the model. Figure 6.3 shows the total number of label changes per participant during the data collection, with each participant entering an average of 37.35 label changes. The number of samples collected by each user for each class was explored to ensure there wasn't significant bias. Each user successfully collected data for each of the five classes showing similar patterns to the complete dataset, with the majority of users collecting more data for classes 3, 4 and 5. Therefore, the percentage of the data each user collected per class was calculated with an average standard deviation was 0.15 showing that while there is a small class imbalance, no user has a significant class imbalance that would impact the classification model. No single user collected significantly more or less data than the average user with the number of samples collected between users having a standard deviation of 0.01 with the size of each user's dataset ranging from 3.1% to 6.8% of the total dataset.

6.2.2 The Human Activity Recognition Using Smartphones Data Set (WISDM)

Experiment setup: This dataset has been published by Kwapisz JR, Weiss GM, Moore SA (2011) [176]. The dataset comprises of 36 participants performing six physical activities (walking, walking upstairs, walking downstairs, sitting, standing and jogging) while carrying the provided Android-based smartphone in their front trouser leg pocket. The 36 participants resulted in 1,098,207 data samples being collected consisting of 39% walking, 31% jogging, 11% walking upstairs, 9% walking downstairs 6% sitting and 4% standing.

Sensors: The dataset comprises of accelerometer data from the embedded sensor within the Android phone sampled at 20Hz. A graphical user interface was included within the custom Android application enabling users to label the 6 prespecified physical activities (walking, jogging, walking upstairs, walking downstairs, sitting and standing) when commencing a new activity and to end the data collection.

This external dataset is freely available at:
http://www.cis.fordham.edu/wisdm/dataset.php

6.3 Methods

Using the three aforementioned datasets, two TL approaches are proposed to help improve accuracy and reduce the requirement of large datasets when classifying mental wellbeing:

1. Training a base model on a large human activity dataset and transferring the learned knowledge to a 1D CNN trained using the controlled experiment dataset to classify stress.

2. Using signal-image encoding to transform accelerometer data into images and then applying a novel CNN-TL-based approach combined with a separate 1D CNN trained using the remaining sensor data to classify real-world affective state.

6.3.1 Modality Transformation

An image is comprised of pixels which can be conveniently represented in a matrix with a colour image containing three channels; red, green and blue for each pixel, compared with grayscale images that contain only one channel. Transforming time series data into images can help extract multi-level features [336] and improve classification accuracy [357], [157].

This study aims to explore the use of signal-image encoding and the addition of TL with time series data. Therefore the continuous, fast changing datastream of accelerometer data must first be transformed into images. It is not plausible to transform the physiological data into images due to its static nature where HR and EDA can often remain constant for several seconds resulting in no data being encoded. Three methods of modality transformation using accelerometer data are utilised: GADF, GASF and MTF.

Wang and Oates transformed time series data into images using Gramian Angular Field (GAF) [336]. First, the data was normalised between -1 and 1 by applying 6.1. The normalised data is then encoded using the value as the angular cosine and the time stamp as the radius r with 6.2, where ϕ is the angle polar coordinates, ti is the time stamp, N is a constant factor to regularize the span of the polar coordinate system and $\tilde{X}$ represents the re-scaled time series data [357].

$$\tilde{x}^{i}_{-1} = \frac{(x_i - \max(x)) + (x_i - \min(x))}{\max(x) - \min(x)} \tag{6.1}$$

$$\begin{cases} \phi = \arccos\left(\tilde{x}_i\right), -1 \leq \tilde{x}_i \leq 1, \tilde{x}_i \in \tilde{X} \\ r = \frac{t_i}{N}, t_i \in \mathbb{N} \end{cases} \tag{6.2}$$

$$GASF = [\cos(\emptyset_i + \emptyset_j)] \tag{6.3}$$

$$= \tilde{X}' \cdot \tilde{X} - \sqrt{I - \tilde{X}^2}' \cdot \sqrt{I - \tilde{X}^2} \tag{6.4}$$

$$GADF = [\sin(\emptyset_i - \emptyset_j)] \tag{6.5}$$

$$= \sqrt{I - \tilde{X}^2}' \cdot \tilde{X} - \tilde{X}' \cdot \sqrt{I - \tilde{X}^2} \tag{6.6}$$

The normalized data is then transformed into polar coordinates instead of the typical Cartesian coordinates. After transformation, the vectors are transformed into a symmetric matrix called the Gramian Matrix. There are two ways to transform the vectors into a symmetric matrix: GASF and GADF as shown from 6.3 to 6.6 where $\emptyset$ is the angle polar coordinates. These methods preserve the temporal dependency, with the position moving from top-left to bottom-right with time.

$$M = \begin{bmatrix} w_{ij|x_1 \in q_i, x_1 \in q_j} & \cdots & w_{ij|x_1 \in q_i, x_n \in q_j} \\ w_{ij|x_2 \in q_i, x_1 \in q_j} & \cdots & w_{ij|x_2 \in q_i, x_n \in q_j} \\ \vdots & \ddots & \vdots \\ w_{ij|x_n \in q_i, x_1 \in q_j} & \cdots & w_{ij|x_n \in q_i, x_n \in q_j} \end{bmatrix} \tag{6.7}$$

Alternatively, images can be generated using MTF where the Markov matrix is built and the dynamic transition probability is encoded in a quasi-Gramian matrix as defined in 6.7. Given a time series x and its q quantile bins each x_i is assigned to the corresponding bins, q_j ($j \in [1, q]$). A q X q Markov transition matrix (w) is created by dividing the data into q quantile bins. The quantile bins that contain the data at time stamp i and j (temporal axis) are q_i and q_j. The information of the inter-relationship is preserved by extracting the Markov transition probabilities to encode dynamic transitional fields in a sequence of actions [336]. A comparison of identical X, Y, Z and total acceleration data transformed as GASF, GADF and MTF can be seen in figure 6.4.

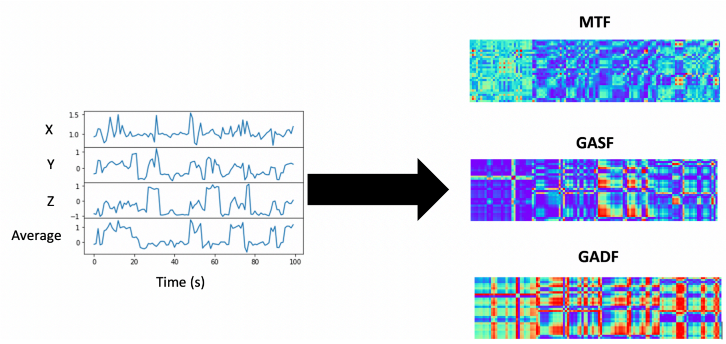


Figure 6.4: An example of raw accelerometer data (X, Y, Z and average motion) transformed using MTF, GASF and GADF.

6.3.2 Model Architecture

Two TL approaches are explored to classify the mental wellbeing data collected from the controlled stressor experiment and the EnvBodySens datasets.

6.3.2.1 1-Dimensional CNN TL Model

The first TL approach aims to improve the accuracy of stress recognition using the controlled experiment dataset. A source 1D CNN is first trained using the accelerometer data from the WISDM dataset for activity recognition and then a TL approach is applied to transfer the learned general characteristics of time series data towards the smaller controlled stressor dataset. The model is trained over 10 epochs with a batch size of 128. Batch normalisation has been used within the network to normalise the inputs of each layer and a dropout layer with a rate of 0.5 was added before the maxpooling layer to prevent overfitting. The pooling layers subsample the data, reducing the number of weights within that activation. Finally, the fully-connected layers where each neuron is connected to all the neurons in the previous layer are used to calculate class predictions from the activation.

When training the physiological model using TL, the source model is imported without its last layer with dense layers then added to enable the new model to learn more complex functions from the HR (BPM), raw HR signal (amplitude), HRV and EDA data. The dense layers used the ReLU activation function and the final layer, which contains two neurons, one for each class (stressed and relaxed) used the Softmax activation function. Hold-out validation is used to test this approach by using a randomly selected 20% test split to test the model and calculate accuracy. This approach aims to reduce overfitting and improve the accuracy in which stress can be classified from a limited sample by transferring the weights from the source model.

6.3.2.2 Image Encoding TL Model

This second approach proposes the novel combination of a 2D CNN utilising TL with signal-encoded images and a 1D CNN model to improve the accuracy of mental wellbeing classification from the EnvBodySens dataset. TL first requires a pre-trained model; for this work multiple pre-trained object recognition networks have been explored. Given that the majority of pre-trained models for TL have been trained on images it is beneficial to train these networks using signal encoded images from the continually changing motion data and not the physiological data which can often remain static resulting in little data being encoded.

This approach transforms the accelerometer data from the EnvBodySens dataset into images using GADF, GASF and MTF, resulting in a total of 17,750 images for each encoding technique. These images are then used as input (size 64 X 64 X 3) to train a 2D CNN consisting of 2 convolutional layers, polling layer, dropout layer and fully connected layer over 10 epochs to classify 5 states of well-being by exploring 7 pre-trained models (Xception, VGG19, ResNet, NasNet, DenseNet, DenseNet V2 MobileNet) to apply the TL approach. An overlapping sliding window strategy has been adopted to segment the motion data with a window size of 100 samples and a step of 20 chosen experimentally, resulting in an embedded vector of size 4X4 being fed into the fully connected layers. An additional 1D CNN model using the same parameters as the aforementioned 1D CNN is trained using the remaining sensor data from the EnvBodySens dataset (HR, EDA, body temperature, acceleration, noise, UV and air pressure). Batch normalisation and dropout layers with a rate of 0.5 have also been utilised. These two models are frozen and then the concatenated feature vector is fed into two fully-connected layers as shown in figure 6.5.

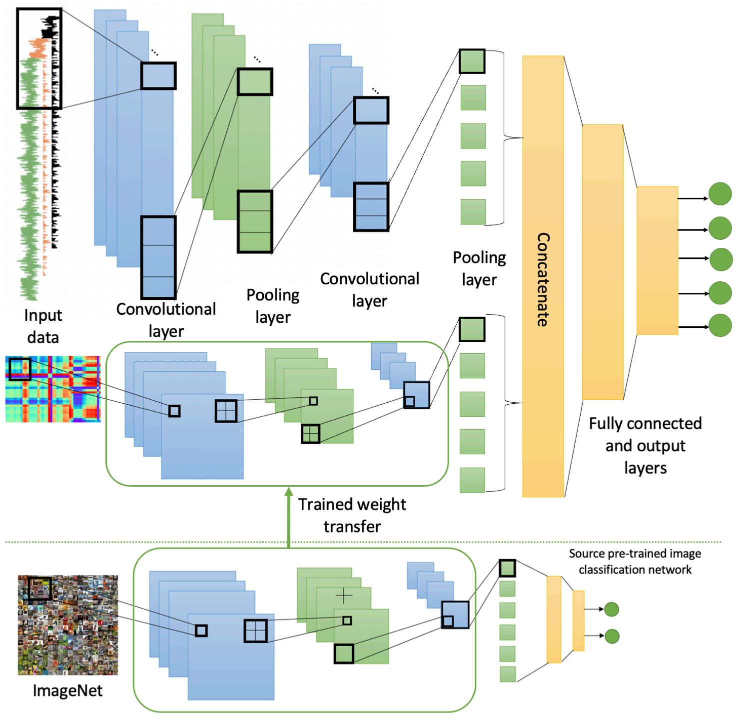


Figure 6.5: Combinatory model consisting of 1D CNN trained using raw physiological sensor data (top) and a 2D CNN using a transfer learning approach trained using accelerometer encoded images (bottom).

The use of TL to transfer weights from pre-trained image classification models and the use of dropout layers help to prevent overfitting, where a model fits exactly against its training data and is unable to classify new data. The use of methods to prevent overfitting is necessary as while the EnvBodySens includes data from 20 participants, deep learning models typically require vast datasets to accurately classify without overfitting.

Hold-validation using a 20% test split has been used to test the model using around 284,000 sensor data samples for training and 71,000 for testing. Additionally, Leave-One-participant-Out Cross-Validation (LOOCV) has also been utilised to test the TL approach on a subject-independent basis. This is where the model is trained with 19 users' data then tested on the remaining user's data (19926 average data samples) to better simulate how the model would be used in the real-world to infer an individual's wellbeing.

6.4 Results

6.4.1 1D CNN TL Stress Model - WISDM & Controlled Stressor Data

As the controlled stressor experiment only collected physiological data (HR BPM, raw HR amplitude, HRV, EDA) it is not possible to use the same modality transformation techniques to transform fast changing motion data as used with the EnvBodySens dataset. However, it is possible to perform TL by initially training a similar model and then using a TL approach to adapt the model with the target stressor dataset.

Initially a 1D CNN model was trained using the physiological data from the controlled experiment dataset, achieving a baseline accuracy of 82.5%. To evaluate whether TL can improve model performance the same network architecture was trained using the larger WISDM dataset consisting of accelerometer data for six human activities, achieving 91% accuracy.

The activity recognition model was then used to apply the TL approach. Once the model had been re-trained with the target controlled stressor dataset an accuracy of 91% was achieved using a 20% test split, an 8.5% accuracy improvement

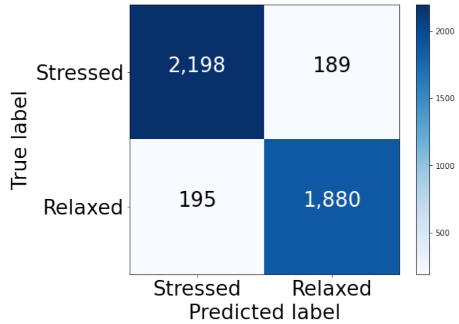


Figure 6.6: Confusion matrix of 1D CNN using TL to infer stressed and relaxed states of wellbeing from the controlled stressor dataset.

over the non-TL 1D CNN approach as shown in figure 6.6. This demonstrates the potential of TL to have a significant impact in improving the classification accuracy of stress recognition without the need to collect extremely large datasets.

6.4.2 Image Encoding Transfer Learning Model - EnvBodySens

The second approach used the EnvBodySens dataset to explore the multi-class problem of classifying five emotional states using the signal-image encoding TL model. Seven pre-trained models (Xception, VGG19, ResNet, NasNet, DenseNet, DenseNet V2 MobileNet) were used to explore the TL approach for the three methods of signal-image transformation (GADF, GASF MTF). Each of the approaches used transformed the motion data from the EnvBodySens dataset to images using a TL approach paired with a 1D CNN to train the remaining time series data from the EnvBodySens dataset. The final testing accuracy using the 20% test data split are reported for each model in Table 6.2.

6.4.2.1 Comparison of Data Modalities

The data modalities were investigated to explore which modalities most contributed towards the classification of mental wellbeing. When all sensor data (HR, EDA, UV, body temperature, air pressure and noise) was used to train the 1D CNN combined with the TL approach for the signal-image transformed

Table 6.2: Comparison of accuracy for different pre-trained deep learning models adapted for mental wellbeing classification through TL trained using only physiological data (HR EDA) or all data (HR, EDA, UV, body temperature, air pressure and noise).

	GASF		GADF		MTF	
	Physiological	All	Physiological	All	Physiological	All
Xception	0.975	0.96	0.977	0.971	0.972	0.956
VGG19	**0.984**	0.952	0.98	0.94	0.964	0.95
ResNet	0.963	0.955	**0.978**	0.937	0.964	0.972
NasNet	0.977	0.965	**0.983**	0.963	0.967	0.964
DensetNet	0.975	0.977	**0.985**	0.971	0.97	0.977
MobileNetV2	**0.981**	0.97	**0.981**	0.954	0.979	0.97
MobileNet	**0.98**	0.967	0.968	0.955	0.974	0.959
No TL	**0.98**	0.974	0.974	0.963	0.975	0.968

motion data, the model classified the 5 emotional states with accuracies between 93.7% and 97.7%. The 1D CNN was also trained using only physiological data (HR EDA) to examine the impact not including environmental data has on model performance. When using the TL approach for motion data and a 1D CNN trained using only physiological data, the model accuracy increased to the highest achieved accuracy of 98.5% when using GADF to transform the motion data and DenseNet to perform TL, as shown in figure 6.7. Furthermore, when comparing the highest accuracy for each pre-trained CNN the physiological model consistently outperformed the model trained using all modalities. This demonstrates the importance of physiological data when determining wellbeing state, unlike environmental data which resulted in more misclassification errors, in particular class 5 the poorest mental wellbeing state.

To evaluate whether the signal-image encoding TL approach improves model performance all sensor data (HR, EDA, noise, UV, body temperature, air pressure and accelerometer data) was used to train the 1D CNN model without performing TL or signal-image encoding. The 1D CNN achieved 93% accuracy, an overall reduction in accuracy compared with the TL model. Furthermore, when the same 1D CNN was trained again using only physiological data, the model achieved 94% accuracy, as shown in figure 6.8, a 4.5% reduction in accuracy. This demonstrates that image encoding and TL can increase overall model accuracy but performance

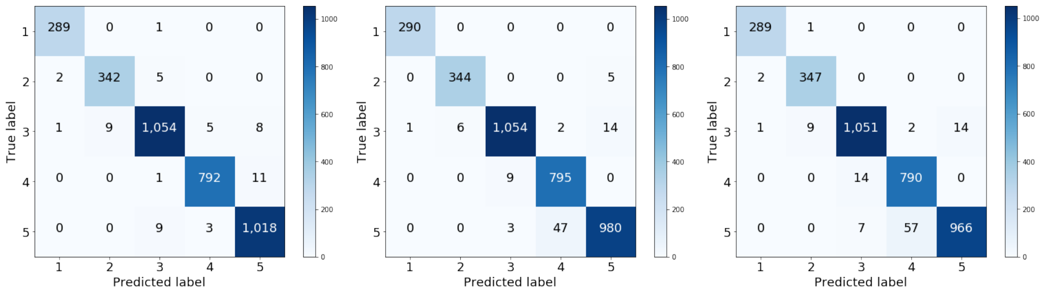


Figure 6.7: Confusion matrix for DenseNet model trained using HR, EDA and GADF (left), GASF (middle) and MTF (right) encoded motion data.

is highly dependent on the additional sensor modalities used to train the network.

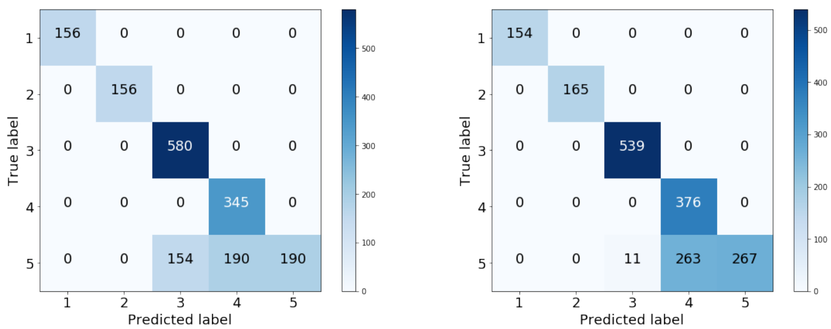


Figure 6.8: Confusion matrix of CNN model without TL when trained using all data achieving 93% accuracy (left) and only physiological and motion data achieving 94% accuracy (right).

6.4.2.2 Comparison of Pre-trained Models

To explore whether the high accuracy achieved was influenced by the pre-trained model used in the TL approach, other pre-trained CNNs were tested using the same GASF, GADF and MTF transformed images. As shown in table 6.2 DensNet achieved the highest accuracy for the GADF transformed data although VGG19 achieved the highest accuracy for GASF data and MobileNetV2 for MTF data. This demonstrates that the pre-trained model selected has little impact on performance with the average variance between the best and worst performing model for all 3 image encoding techniques being only 1.77% for the physiological models and 2.87% for the models trained using all sensor data.

6.4.2.3 Comparison of Signal-Image Encoding Techniques

The signal-image encoding technique also impacted model performance. GASF and GADF outperformed MTF for each pre-trained model, where GADF achieved the highest performance for four of the pre-trained models and GASF for the remaining three. The average accuracy for the GADF physiological model was 97.9% compared with 97.6% for GASF and 97% for MTF showing negligible variations in performance between the different techniques.

6.4.2.4 Subject-Independent Models

As the GADF signal-encoding technique slightly outperformed the other encoders, it was used to explore subject-independent physiological models. Table 6.3 shows the accuracy achieved for each of the 20 users when the model was tested using LOOCV with each individual's physiological data. The accuracies range between 36.4% for user 1 and 77.7% for users 16 and 17. The outlier low accuracy for user 1 is due to corrupt EDA data which continually recorded null readings. The remaining users demonstrate more consistent accuracies and while lower than when tested using hold-out validation, they demonstrate the possibility of inferring wellbeing on an individual basis.

Table 6.3: Comparison of subject-independent classification accuracy for 20 users using the TL approach with GADF encoded images.

User	Accuracy	User	Accuracy
1	0.364	11	0.709
2	0.698	12	0.734
3	0.702	13	0.723
4	0.683	14	0.738
5	0.666	15	0.749
6	0.752	16	0.777
7	0.736	17	0.777
8	0.706	18	0.753
9	0.737	19	0.690
10	0.636	20	0.763

The subject-independent models were also trained without the TL approach while still transforming signals into images to explore whether performance improvements were due to TL. A 2D CNN was implemented to train the signal encoded images which was concatenated with the 1D CNN trained using the physiological data. The results show TL increased average accuracy by 0.55% for all users which falls within the margin of error, demonstrating no overall performance improvement. However, the TL approach never degraded the performance of individuals' models and achieved up to a 4% increase in accuracy showing it can be beneficial and should continue to be used.

6.5 Discussion

A new CNN-TL-based approach towards affective state classification has been introduced that goes beyond previous signal-image encoding frameworks by incorporating TL in addition to a separate 1D CNN. This research demonstrates that a signal-image encoding TL approach can improve the performance in which five affective states can be classified, achieving up to 98.5% accuracy using hold-out validation and an average of 72.3% using LOOCV. This outperforms many previous real-world affect recognition systems [55], [214], [138], [363] including a previous stacked machine learning approach using the same EnvBodySens dataset which achieved an accuracy of 86% [156] and a combined CNN and RNN using the same dataset that achieved 94.9% accuracy [155]. Furthermore, a TL approach to alleviate limited wellbeing data by transferring knowledge from a model trained using a large physical activity dataset improved accuracy to 91% when classifying stress.

The results have demonstrated that the integration of TL as part of the newly proposed methodology, extending standard deep learning algorithms can greatly improve the classification of affective state. In particular, the combinatory approach of encoding accelerometer data as images using GADF, GASF and MTF then using a pre-trained model to perform TL and combining this model with a 1D CNN trained using physiological data, has improved the accuracy with which affect can be classified. The encoding technique used was shown to have a minor impact in model accuracy demonstrating GADF was most effective for the majority of the models. Similarly, the pre-trained model used to perform TL had a limited impact on model performance with an average difference of only 2.32% between the different models. However, TL only slightly improved performance by an average of 0.55% when testing using subject-independent models, demonstrating the transformed images had a greater impact on model performance than TL.

Furthermore, solely using physiological and motion data resulted in the highest accuracy (98.5%), outperforming models additionally trained using environmental data. This suggests that environmental factors such as noise and UV are more challenging to use for affect recognition even when paired with physiological data. The reduced performance may be due to the intricate information in the environmental data already being captured inherently in the physiological and motion data for example poor weather having a negative impact on mood.

When testing using LOOCV the subject-independent accuracies are lower than subject-dependent accuracies. The average accuracy of the subject-independent physiological models excluding user 1 was 72.3% (SD 0.038), compared with 98.5% for the subject-dependent model both using GADF to transform the signals and a DenseNet pre-trained model. This likely reflects that different individuals have different patterns of physiology when experiencing the same state of wellbeing and that similar levels of activity are perceived differently in terms of valence [314] demonstrating similar results as other studies [148], [190], [13].

An additional TL approach involved training a 1D CNN using a human activity dataset and then performing TL to adapt the model using physiological data from a controlled stressor experiment. This TL approach improved model accuracy by 8.5% over the source model achieving 91% accuracy, again demonstrating the benefits of TL approaches. However, this model did not perform as well as the image encoded TL model even though it was only classifying two classes in controlled conditions compared with five states of real-world affect. This demonstrates the vital role motion data encoded as images plays in improving wellbeing classification, although it may not always be possible to collect motion data such as during the controlled stressor experiment where participants were stationary.

Overall, three multivariate datasets were used as benchmark datasets to evaluate the TL approaches. This work demonstrates that by using the proposed approaches it is possible to capitalize on two modalities to accurately classify wellbeing on a 5-point Likert scale. The results have demonstrated that TL approaches are appropriate for modeling affective states especially when training data is scarce. These TL approaches have outperformed previous models using the same EnvBodySens dataset built on ad-hoc extracted features [156] and 2 dimensional CNNs [155]. These findings showcase the potential for TL and signal-encoded images to improve affective multimodal modeling.

6.6 Conclusion

Recent developments in tangible interfaces and edge computing are producing sensory datasets as people are going about their daily activities. However, accurately classifying these limited datasets can be a challenging proposition. In this chapter, a scenario of wellbeing classification using small multimodal datasets has been presented. Although these types of time series datasets can help us understand people's wellbeing, current recognition techniques are not efficient enough to tackle data scantiness. TL offers an automated way to utilise the learning outcomes from larger datasets to help overcome these challenges. This research has demonstrated the advantages of employing a combinatory TL approach for raw multimodal sensor data modelling.

Based on experimental results, the proposed frameworks employ two TL models using three multimodal sensor datasets. The first approach trained a 1D CNN model using the WISDM dataset and performed a TL approach to adapt the model for the target stress dataset of physiological signals. This model achieved 91% accuracy, a 8.5% improvement over the same 1D CNN trained without TL. The second model used the proposed framework combining a TL approach using signal encoded images with a 1D CNN. Accelerometer data was transformed into RGB images using GADF, MTF and GASF. This data was subsequently used to train a 2D CNN using pre-trained models to apply a TL approach. This model was concatenated with a 1D CNN architecture trained using raw physiological sensor data and resulted in increased performance, achieving up to 98.5% accuracy. Overall, incorporating a TL signal-image encoding approach in addition to a 1D CNN helped further improve model performance with limited data.

Chapter 7

Research Applications and Real-Time Interventions

This chapter explores and presents a number of applications of this research that pursue the design of tangible fidgeting tools as enjoyable, non-intrusive interfaces for monitoring and improving mental wellbeing. In particular, three potential applications are considered for two target groups, these are: 1) iFidgetCube [347] - a graspable device that can help tackle stress and anxiety for the general population, 2) Fidget watch - an adapted version of tangible fidgeting interfaces that leverages wearable technology for ease of use and portability and 3) Tang-Toys [349] - smart toys that can communicate and support children. This section is adapted from [347], previously published in IEEE Sensors Journal, [349], previously published in Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computer and [351] previously published in UbiComp/ISWC 2018 - 2018 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2018 ACM International Symposium on Wearable Computers.

7.1 Introduction

Limited attention has been given to the design of technological solutions for individuals who might benefit from self-support wellbeing tools. Existing technologies commonly range from online therapy programs (e.g., Computerised Cognitive Behavioural Therapy (CBT)) for depression [79]) and self-help systems, to designs that supplement psychotherapy by providing additional content to support mindfulness [69] and remote monitoring [226]. These systems are limited to the monitoring of clients using sensors while relying on users to act, be able to recognise and verbalise, or self-report accurate information in relation to their health and physiological status. Users often do not, or may not, know how to respond, therefore automatically providing specific direct physical feedback may ameliorate this effect. In addition, the ability to access vulnerable and underserved groups have made it possible to design more effective interventions that can be tailored to their needs, where technology can have a life-changing impact.

When stressed, people commonly fidget with physical objects such as pens. TUIs can help promote fidgeting as it is a natural response that demonstrates the potential to regulate stress [96], [211], provide a distraction which can significantly reduce anxiety [141] and improve information retention [90]. Recently, fidgeting cubes have begun to increase in popularity; they are small plastic cubes whose sides provide sensory tools to facilitate fidgeting and help normalise stimming (self-stimulatory behavior such as tapping or clicking). Previous work has found that within fidgeting tools most users preferred squeezing (89%) and stretching (79%), but this was closely followed by clicking, pressing and tapping (71%) which could easily be embedded within TUIs [65]. Sensory tools embedded within interfaces may assist people suffering from a range of mental health conditions to relax and provide a distraction, as 79% of children who fidgeted did so to regulate their emotions [65].

Tangible Fidgeting Interfaces (TFIs) have been introduced as physical fidgeting devices that enable repetitive physical interaction while also enabling objective sensor measurement. By including fidgeting mechanisms within TUIs they can act as a distraction to any wellbeing challenges encountered, which is often used as a coping strategy to reduce stress [53], [151]. These fidgeting interfaces can also be coupled with the developed deep learning algorithms to infer and improve real-world mental wellbeing. TFIs can provide aesthetically pleasing user interfaces in any form, enabling them to ubiquitously become part of everyday interactions. They can be of any shape (e.g. cube, ball or polytope) and can be made of hard or soft material (squeezable ball or clicking tool). These digitally enabled fidgeting mechanisms offer a variety of sensory actions catering for a wide range of needs in a small, unobtrusive design providing a distraction and unmet need for people experiencing anxiety or stress.

By developing handheld TFIs that embed physiological sensors, it is possible to develop monitoring devices that encourage engagement and improve wellbeing. Therefore, three research applications have been developed taking advantage of physical interactions with TUIs to promote fidgeting, automatically apply calming haptic feedback in real-time and aid communication between children between through play.

7.2 Applications

7.2.1 iFidgetCube

The first application of this research is the development of iFidgetCube, a TFI in the shape of a cube whose various sides can provide multiple sensory tools to interact with, similar to traditional fidget cubes. This small plastic cube provides sensory tools to interact with such as buttons, as shown in figure 7.1. Unlike traditional fidgeting cubes, iFidgetCube embeds a microcontroller and non-invasive sensors to digitise and enable the real-time monitoring of physiology and fidgeting interactions. This presents the first device combining traditional fidgeting cubes with a microcontroller and non-invasive physiological sensors.

Multiple physiological sensors are included within the interfaces to monitor real-time changes. Physiological sensors measuring HR, HRV EDA are embedded on opposite faces of the cube, allowing two fingers to be placed to comfortably hold the device while simultaneously recording sensor data.

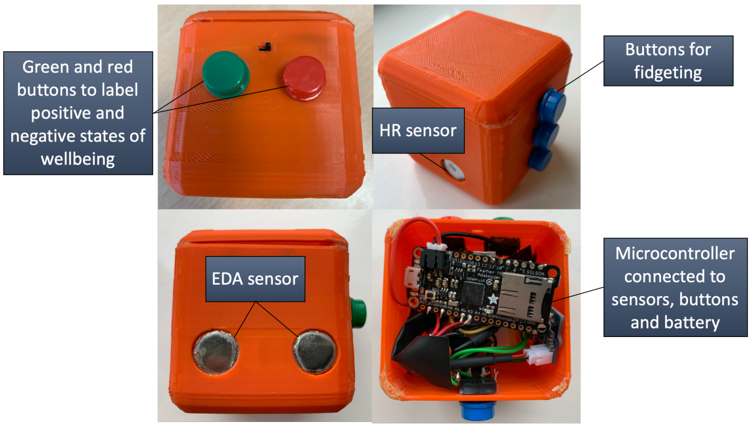


Figure 7.1: iFidgetcube showing labelling buttons, fidgeting buttons, HR sensor and EDA sensor.

The iFidgetCube contains 3 buttons that can be used for fidgeting similar to fidgeting toys along with a 9-DOF IMU to capture the fidgeting motion of interactions. The IMU is a MPU-9265 and consists of an accelerometer, a gyroscope, and a magnetometer operated at 3.3v, footprint of 22 $\times$ 17 mm and connected to a ATmega32u4 based processor to process all of the data with a footprint of 28.8 $\times$ 33.1 mm, powered by a small size 3.7v lithium polymer battery as shown in figure 7.2.

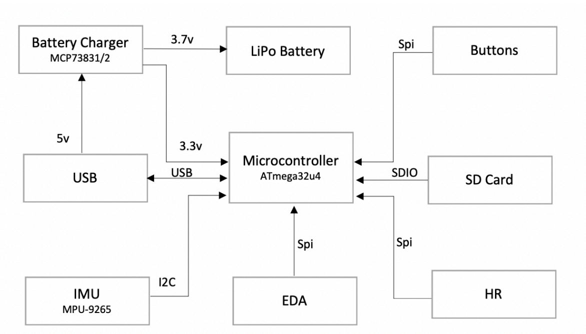


Figure 7.2: TFI electronic schematic.

7.2.1.1 Evaluation

During the real-world data collection discussed in Section 4.3, 4 participants experienced the fidgeting cubes. During a focus group participants stated they enjoyed using the physical interfaces, finding them easy to handle and engaging unlike other physiological sensing devices. The interfaces were described as *"calming"* and *"relaxing"* due to being able to fidget by moving the devices as well as pressing the buttons.

Real-world trials with the devices enabled the collection of labelled physiological sensor data used to train the developed classification models. During the data collection period, users consistently moved the interfaces and used the fidgeting buttons, demonstrating their simplicity and effectiveness. Users believed using the fidgeting buttons helped them relax and improved their mental wellbeing however due to the nature of the device it was not always possible to use the physiological sensors while fidgeting.

7.2.2 Fidget Watch

The second research application extends the use of fidgeting tools by additionally exploring the feasibility of wellbeing classification models activating instant haptic feedback upon changes in wellbeing, in the form of subtle vibrations experienced as a tap on the wrist.

A watch has been developed that embeds physiological sensors measuring HR, HRV and EDA in addition to incorporating fidgeting buttons similar to the iFidgetCube. The Fidget watch is smaller and easier to use during daily activities than iFidgetCube and also includes haptic feedback in the form of vibrating motors that are small in size and require very low power to operate.

The Fidget watch incorporates Bluetooth to connect to a custom built Android application. Due to the resource constrained nature of the device it is not capable of running the developed classification models on-device. Instead, the physiological data is continuously transmitted wirelessly to the Android app that is capable of running in the background without requiring any action from the user to ensure the data is always received. The app inputs the physiological data into the developed personalised classification model and when poor mental wellbeing is inferred automatically activates calming haptic feedback.

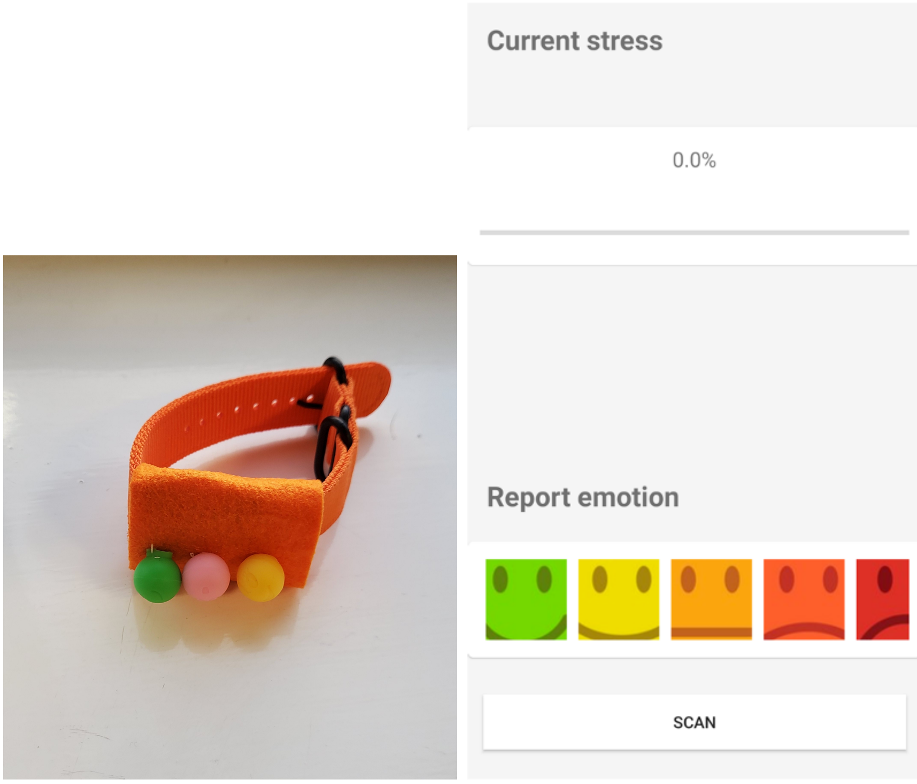


Figure 7.3: Example of a Fidget watch (left) and connected mobile application (right).

Utilising a mobile application to run the classification model helps the wearable interface remain small by reducing the processing power required. The application allows the user to wirelessly pair their Fidget watch, enabling the continuous transmission of all physiological data in the background. The application also enables users to view the current classification output on a scale from poor mental wellbeing to positive mental wellbeing as well as the percentage predicted by the classification model, as shown in Figure 7.3. Once poor wellbeing is inferred the haptic feedback is actuated at a frequency of 20% lower than the participant's HR [159] between 40BPM and 65BPM [63] as used in previous studies to simulate a subjective state of calmness [64]. The personalised haptic feedback aims to passively calm users whilst also actively alerting users of their current wellbeing allowing them to reflect.

The Fidget watch has been trialled in laboratory settings to test its functionality where it was found to continuously monitor wellbeing due to its on-body placement and allowed for easy access to the fidgeting buttons. Overall, this device combines the real-world application of mental affective state classification with the automatic delivery of tangible interventions to improve wellbeing.

7.2.3 Tangible Toys

Children's toys represent another ideal embodiment for TUIs as they provide sufficient space for the electronics and encourage tactile interactions. Although a limited number of TUIs for mental wellbeing have previously been developed by researchers, many of these were not engaging for children and often contained physiological sensors which prevented physical interactions commonly used by children to interact with objects such as toys. An interface that can actively monitor and enable the communication of a user's physical interactions and wellbeing would be beneficial for all.

Few sensor based interfaces have been designed for children even though they traditionally find it challenging to communicate their mental wellbeing [246]. The ability for children to communicate their wellbeing is vital as children with difficulties communicating are more at risk in terms of social acceptance and bullying [191]. Furthermore, better relationships and communication with friends offers protection against poor mental health in the future [92]. This research application introduces Tangible Toys (TangToys) with the aim of enabling children to communicate wellbeing through embedded sensors and feedback actuators. The devices embed sensors used to measure physical interactions and mental wellbeing along with Bluetooth Low Energy (BLE) to enable real-time communication. Through the use of BLE TUIs can communicate with one another, enabling real-time communication networks to be developed. The ability for devices to communicate with each other enables children to communicate when socially distant and the ability to discover other nearby users.

TangToys can vary in shape, size and material. The four TangToys developed have been designed for younger children aged 5-7 as this is when children develop self-conscious emotions and develop an emotional front [260], they include 2 soft teddies, a soft ball and a cushion. As children physically interact with TangToys in the same way as traditional toys all of the interfaces are suitable for children and encourage engagement by resembling familiar toys.

Each TangToy includes a microcontroller and micro SD card to record all interactions along with BLE 4.2 for communication. A range of sensors can be used to monitor children's interactions with the toys including capacitive sensors to measure touch and 9-DOF IMU to measure motion. Physiological sensors were not embedded within the toys as younger children may not understand how to operate the sensors and they may hinder physical interactions whilst playing with the toy.

In addition to the sensors, TangToys can provide real-time feedback activated by wireless communication with other TangToys. Haptic feedback has been included within some of the developed prototypes issuing a physical sense, resembling touch and providing comfort which can improve mental wellbeing [62], [17]. Additionally, visual feedback in the form of multi-coloured LEDs has been included within the soft ball and teddy prototypes.

7.2.3.1 Communication Framework

Embedding sensors within toys that can communicate with one another through BLE offers many new opportunities for real-time interactions. BLE 4.2 has a range of around 50m allowing TangToys to communicate with one another in locations such as playgrounds. Two opportunities are presented for real-time digital social interaction between TangToys.

By utilising Peer to Peer (P2P) communication it is possible for two connected devices to directly communicate with each other. This method of communication helps friends who may be nearby but socially distanced to provide physical communication that is not possible with other devices. When a child plays with a TangToy the capacitive sensor and accelerometer data measuring touch and motion respectively can be actuated on the paired interface through the embedded haptic and visual feedback to simulate physical communication. The connected friend can then react to this communication by interacting with their device allowing friends to wirelessly support one another through physical interactions.

Table 7.1: Encoded haptic patterns for different states of wellbeing to be activated on connected TangToys based on touch and motion interactions.

Variable	Positive wellbeing	Neutral wellbeing	Negative wellbeing
Touch	Low to Med intensity, no sudden changes	Irregular fluctuations. Inconsistent frequency changes and amplitude	Fluctuations more pronounced returns to shifting baseline erratically
Motion	Low activity. Activity fairly smooth	Consistently Low to Med level of fluctuations and frequency of changes. Amplitude mostly Low to Med	High activity. Occasionally in bursts. Baseline is irregular
Haptic pattern	Low frequency, low intensity	Low to Med frequency	High frequency, high intensity

The range of feedback offered differs depending on the interactions with the paired devices as shown in Table 7.1. For example, if a child is aggressively shaking their TangToy or touching it harshly this can result in prolonged sharp haptic feedback patterns being played on the paired device and red visual feedback. This enables friends to physically communicate how they are feeling and provide comfort to one another by replying with soft, gentle interactions to provide comfort and a sense of presence, as shown in Figure 7.4.

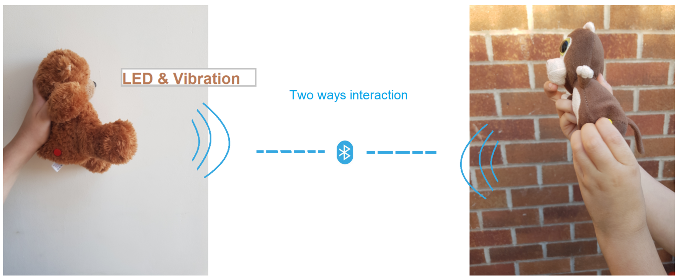


Figure 7.4: Two children playing using TangToys.

Each TangToy can also use its Bluetooth capabilities to broadcast its presence to other TangToys. When a TangToy detects another device nearby this can initiate feedback being issued to alert the child of other nearby children. This allows a child to find other children who may require support when not near their friends to facilitate P2P communication. These children can then interact with the devices to form a support group to communicate their wellbeing to each other. The feedback actuated when detecting other devices can be impacted by the number of nearby interfaces. For example, if a single child is detected nearby, more subtle haptic feedback is issued compared with more pronounced feedback when multiple TangToys are nearby. Similarly, the colour displayed on the TangToy can change dependent on the number of users located nearby to alert the user visually. Using this method of interaction does not enable the same capabilities as the P2P communication, but enables each device to interact automatically with other nearby devices, and afford a sense of *'togetherness'*.

7.2.3.2 Empirical Evaluation

TangToys have been presented in focus groups to members of the NICER group and teachers to provide feedback on the design and functionality of the interfaces. Teachers considered the methods used to interact with TangToys suitable for children and believed the way in which children interact with the toys will likely indicate their wellbeing. Additionally, teachers liked the design of the toys as they appear similar to other toys helping to reduce stigma. Overall, the participants reported the design, sensors and communication capabilities were all suitable for children and believed TangToys would promote the communication of wellbeing between friends.

7.3 Conclusion

Overall, three real-world applications have been developed to monitor and improve mental wellbeing. iFidgetCube provides a simple method to both monitor physiological changes and simultaneously provide a fidgeting interface as a preventative mechanism to aid relaxation and ease restlessness. The Fidget watch incorporates physiological sensors and fidgeting mechanisms within a wearable interface utilising personalised models to automatically activate haptic feedback when poor wellbeing is inferred. Finally, TangToys present a new opportunity to monitor children's physical interactions which may help provide an indication of their mental wellbeing. The ability for TangToys to communicate with one another provides a non-intrusive means for children to communicate their wellbeing through play.

These research applications demonstrate the potential for tangible interfaces to monitor real-world affective state and actuate interventions in real-time. In particular, the Fidget watch demonstrates the potential to classify real-world physiological data and automatically activate real-time interventions when required. Overall, these applications including fidgeting mechanisms and haptic feedback help to automatically improve real-world mental wellbeing and demonstrate the benefits of tangible interfaces, the developed classification models and real-time feedback.

Chapter 8

Conclusion and Future Work

This thesis has presented the design of novel TUIs and the development of classification models to advance the understanding and measurement of affective state in real-world environments. This final chapter provides a general discussion of the work, a summary of the different contributions, potential areas of improvement and future work.

8.1 Conclusion

This thesis has contributed to the development of methods to improve the classification of real-world affective state using non-invasive physiological sensors embodied within co-designed tangible interfaces. In particular, this thesis has studied how TUIs can help advance the real-world measurement of affect by exploring:

1. The co-design of TUIs with people who have intellectual disabilities that go beyond existing devices to offer continuous monitoring and feedback mechanisms.

2. The exploration and use of tangible labelling methods to enable the collection of in-situ physiological labelled data which is notoriously challenging to collect.

3. The implementation of two transfer learning approaches to personalise affective models on-device and reduce the necessity for large real-world

datasets to accurately train deep learning classifiers.

4. The development of three research applications acting as preventative interventions to improve wellbeing by providing calming sensory tools, issuing real-time haptic feedback and aiding the communication of children's wellbeing.

To investigate the use of TUIs for mental wellbeing, several explorations, experiments, and real-life studies have been conducted. Based on these efforts, this section summarises the main findings.

Participants with intellectual disabilities often have their wellbeing challenges misattributed to their disability [102]. To help develop solutions relevant to them and the wider population, they were invited to help co-design TUIs for affective state recognition as existing devices cannot continuously, accurately and unobtrusively measure physiological changes in-situ. Chapter 3 explores the methods developed within a series of co-design workshops to engage people with intellectual disabilities in the development of TUIs, including the use of interactive sessions to explore the design process in addition to the sensors and feedback mechanisms used within the interfaces. The participants' contributions resulted in the development of numerous tangible interfaces including 3D-printed interfaces for older children and adults embedding physiological sensors and soft toy devices for children measuring touch and motion interactions.

Before the developed interfaces could be leveraged for real-world data collection, methods to gather accurate self-reported data in real-world environments were required. This is vital for improving our understanding of affective state due to the inability to label sensor data after the point of collection. The developed LabelSens framework discussed in Chapter 4, explores multiple tangible labelling methods that could be incorporated into the co-designed interfaces to simplify the labelling process. Evaluation of the framework resulted in labelling buttons being embedded within all of the developed interfaces to enable participants to label in real-time. The interfaces were then used by participants to collect real-world labelled affective data.

Choosing an appropriate model to effectively classify real-world affect using physiological signals is a challenging proposition. Physiological signals often consist of time series data with a variation over a long period of time and dependencies within shorter periods. Furthermore, physiology is inherently unique to an individual, especially those with intellectual disabilities, limiting the application of generic models. Therefore, Chapter 5 proposed the development of a TL approach to personalise affective models. Initially, a source CNN was trained using controlled stressed and relaxed data, the model was then adapted using the TL approach to personalise and adapt it for the real-world domain. The results show adopting the TL approach significantly increased model performance with the multivariate physiological and motion affective model achieving an average accuracy of 93.5% compared with the comparative non-TL 1D CNN accuracy of 71.7%. Leveraging motion data helped improve multivariate physiological model accuracy and enabled the development of univariate motion TL models that achieved an average accuracy of 88.1%, demonstrating the importance of motion in real-world affect inference. Furthermore, by utilising advances in edge computing the TL approach was applied on-device. This enabled participants to self-label for only a few days before the model could be personalised to greatly improve accuracy. Empowering sensors with this TL approach in portable interfaces paves the way for continuous real-world monitoring without the need to self-report.

While advances in deep learning are helping advance the accuracy in which affective states can be classified, large datasets of multiple users over long periods of time are usually required to accurately train the models [176]. Therefore, many studies in the domain of affective computing have developed models from controlled experimental data as the collection of real-world labelled sensor data is challenging. Chapter 6 presents a second TL approach that alleviates some of the challenges imposed by deep learning architectures. The signal-image encoding TL approach transformed accelerometer data into images which were used to adapt a pre-trained image classification model, this was then combined with a separate CNN used to train the remaining physiological sensor data. When tested with 20 users self-reporting their emotional wellbeing on a 5-point Likert scale, performance was enhanced, resulting in up to 98.5% accuracy. Subject-independent models using the same approach resulted in an average of 72.3% (SD 0.038) accuracy. This methodology demonstrates improved accuracy and helps alleviate the requirements for large affective datasets with deep learning classifiers.

This research has resulted in the development of many real-world applications, discussed in Chapter 7 that leverage the co-designed tangible interfaces and developed classification models. Children's mental wellbeing is more important than ever before with an increasing number of children experiencing high levels of stress [12]. Therefore, TangToys have been developed to help promote physical communication through play, benefiting those who find it difficult to verbalise their emotions. This technological solution helps children communicate digitally and receive support from one another as advances in networking and sensors have enabled the real-time transmission of physical interactions. Furthermore, iFidgetCubes have been constructed combining physiological sensors with tactile engagement through fidgeting, enabling repetitive interaction methods to tap into an individual's psychological need to feel occupied and engaged. Finally, a Fidget watch has been developed that combines repetitive calming interactions with physiological sensors, enabling automatic inference using a custom mobile app. By developing a wearable and connecting it to an Android application it is possible to continuously and unobtrusively monitor affective state by utilising the developed classification models. If the model infers poor wellbeing, personalised haptic feedback is automatically issued to promote relaxation. These interfaces demonstrate the real-world applications of this research to automatically monitor affective state and provide instant interventions.

Overall, this work has systematically demonstrated that the development of custom TUIs and tangible labelling techniques has aided the collection of real-world wellbeing data. Furthermore, the two TL frameworks proposed demonstrate the ability to personalise affective models and reduce the reliance of large datasets for deep affective modelling. Tangible interfaces and the TL frameworks demonstrate numerous potential applications to monitor real-world affective state and provide therapeutic interventions such as fidgeting tools and real-time haptic feedback. This work has the potential to greatly improve access to tools that assess real-world wellbeing including for those with intellectual disabilities.

8.2 Challenges and Future Work

This thesis presents a step towards real-world affective measurement using the developed tangible interfaces and TL approaches. While the findings of this research are encouraging there are ways in which this work can be extended and improved upon. This section acknowledges the limitations and provides suggestions for future work.

8.2.1 Datasets

While the personalised TL results using the real-world data are encouraging, they are by no means sufficient. This is because the dataset is too small to suggests that the affect recognition systems can generalise to unconstrained wellbeing in the wild. To properly test the boundaries of the proposed systems, they need to be trained on datasets that are large scale and continue to accurately mimic real-life situations.

While the datasets used include physiological sensor data and are representative of a typical affective modelling scenario, the TL approaches could be tested on diverse datasets with more participants and for longer periods of time in various real-world scenarios to ensure their performance is sustained.

8.2.2 Mental Health Recognition

The developed systems monitor affective state but in the future it may be beneficial to further explore the classification of specific mental health conditions. A possible future direction would be to study the detection of anxiety or autism, to help clinicians in medical diagnosis. Currently, the manner in which a patient is diagnosed is based on individual assessment where as the number of patients increases so does the need for accurate diagnosis. Standardisation of the diagnosis task using the developed models could greatly help doctors or counsellors provide personalised care to the most at risk patients.

8.2.3 Model Architecture

The developed TL approaches utilise CNNs as they were shown to outperform the various other models tested. However, with advances in network architectures, new neural networks could continue to be explored utilising the developed TL approaches to further increase affective modelling performance.

8.2.4 Edge Computing

On-device processing presents many future opportunities for this research. Presently, resource-constrained edge computing interfaces such as the Raspberry Pi have been used to develop personalised models. However, the limited processing power of the device limits its capabilities and its large size prevents its use in on-body interfaces. Advances in edge computing may reduce the size of computing devices and remove the reliance of mobile apps, potentially enabling a small wearable interface to perform real-time classification on-device.

8.2.5 Sensors

As part of the analysis, several types of wearable signals have been considered with a focus on the use of non-invasive physiological sensors that can be used unobtrusively in the real-world. In the future, additional contextual and behavioural sensor data could be collected to help better capture different changes. However, the size and invasive nature of the sensors will have to be carefully considered to ensure the devices remain functional in real-world environments.

8.2.6 COVID-19

A limitation of this research is the relatively small sample sizes used during the data collection trials. However, due to the global pandemic of COVID-19 it has not been possible to conduct further experiments involving human participants. In the future, it would be beneficial to trial the developed interfaces and classification models with a larger number of participants. It would also be beneficial to trial the applications of this research including the iFidgetCube, tangible toys and

Fidget watch with additional participants using real-world longitudinal studies to explore the impact of these real-time interventions.

8.3 Summary of Contributions

The research detailed in this thesis makes several original contributions which are summarised below.

8.3.1 Co-designing Tangible User Interfaces

- A range of TUIs to monitor real-world mental wellbeing have been developed with potential end users. The co-design process was adapted to suit the needs of the participants with intellectual disabilities and ensure designer subjectivity was removed. The use of interactive sessions enabled target users to express their decisions and ensured designs were suitable for future users.

- The co-designed interfaces and data collection methodology was evaluated through multiple focus groups, ensuring the suitability to monitor real-world affective state using the developed interfaces and non-invasive physiological sensors.

8.3.2 Real-world Labelled Data Collection

- Tangible methods to label data in real-time have been explored and evaluated as sensor data cannot be labelled after the point of collection. The results demonstrate the benefits of using 2 adjacent buttons to collect reliable and accurate labelled data.

- The tangible labelling methodology and co-designed interfaces enabled the successful collection of a real-world labelled physiological affective dataset that can be used to train classification models.

8.3.3 Deep Transfer Learning Approaches

- Personalising affective models using an on-device TL approach improved individual model performance by an average of 21.8%, including those with intellectual disabilities. The ability to perform this approach on-device using few labelled samples significantly simplified the process of developing personalised models and paves the way for ubiquitous personalised affective modelling.

- Developing an image-encoding TL approach to classify five states of real-world mental wellbeing helped overcome problems with small datasets when training deep learning models. This approach resulted in up to 98.5% accuracy, thus improving the performance of conventional deep learning methods.

8.3.4 Real-time Interventions

- The real-world inference of mental wellbeing provides many opportunities to issue feedback that serves as interventions. Three research applications have been developed including: iFidgetCube - providing a fidgeting interface as a preventative mechanism, TangToys - enabling children to communicate through physical interactions and afford a sense of togetherness and Fidget watch - combining fidgeting tools with real-time classification to activate personalised haptic feedback.

- The ability to run classification models on-device using edge computing or using mobile apps enables automated real-world monitoring using the developed models. Real-world inference has enabled interventions such as haptic feedback to be issued, aiming to improve wellbeing in real-time.

References

[1] Sameera Abdul-Kade and John Woods. Survey on Chatbot Design Techniques in Speech Conversation Systems. *International Journal of Advanced Computer Science and Applications*, 6(7), 2015. 14

[2] Alexander T Adams, Elizabeth L Murnane, Phil Adams, Michael Elfenbein, Pamara F Chang, Shruti Sannon, Geri Gay, and Tanzeem Choudhury. Keppi: A Tangible User Interface for Self-Reporting Pain. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, 2018. doi:10.1145/3173574.3174076. 17

[3] Dawn Adams and Chris Oliver. The expression and assessment of emotions and internal states in individuals with severe or profound intellectual disabilities, 2011. doi:10.1016/j.cpr.2011.01.003. 53

[4] Aino Ahtinen, Elina Mattila, Pasi Välkkynen, Kirsikka Kaipainen, Toni Vanhala, Miikka Ermes, Essi Sairanen, Tero Myllymäki, and Raimo Lappalainen. Mobile mental wellness training for stress management: feasibility and design implications based on a one-month field study. *JMIR mHealth and uHealth*, 1(2):e11, jul 2013. doi:10.2196/mhealth.2596. 16

[5] Lulwah Al-barrak, Eiman Kanjo, and Eman M. G. Younis. NeuroPlace: Categorizing urban places according to mental states. *PLOS ONE*, 12(9):e0183890, sep 2017. doi:10.1371/journal.pone.0183890. 2, 21

[6] Miquel Alfaras, William Primett, Muhammad Umair, Charles Windlin, Pavel Karpashevich, Niaz Chalabianloo, Dionne Bowie, Corina Sas, Pedro Sanches, Kristina Höök, Cem Ersoy, and Hugo Gamboa. Biosensing

and actuation-platforms coupling body input-output modalities for affective technologies. *Sensors (Switzerland)*, 2020. doi:10.3390/s20215968. 60

[7] Miquel Alfaras, Vasiliki Tsaknaki, Pedro Sanches, Charles Windlin, Muhammad Umair, Corina Sas, and Kristina Höök. From Biodata to Somadata. In *Conference on Human Factors in Computing Systems - Proceedings*, 2020. doi:10.1145/3313831.3376684. 42

[8] Salma Alhagry, Aly Aly, and Reda A. Emotion Recognition based on EEG using LSTM Recurrent Neural Network. *International Journal of Advanced Computer Science and Applications*, 8(10), 2017. doi:10.14569/ijacsa.2017.081046. 33, 117

[9] Abdulrahman Alruban, Hind Alobaidi, Nathan Clarke, and Fudong Li. Physical activity recognition by utilising smartphone sensor signals. In *ICPRAM 2019 - Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods*, 2019. doi:10.5220/0007271903420351. 92

[10] Sherin F. Aly and A. Lynn Abbott. Facial Emotion Recognition with Varying Poses and/or Partial Occlusion Using Multi-stage Progressive Transfer Learning. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 11482 LNCS, pages 101–112. Springer Verlag, 2019. doi:10.1007/978-3-030-20205-7_9. 4, 118

[11] American Psychiatric Association. DSM-5 Diagnostic Classification. In *Diagnostic and Statistical Manual of Mental Disorders*. 2013. doi:10.1176/appi.books.9780890425596.x00diagnosticclassification. 13, 52

[12] American Psychological Association. Stress in America Are Teens Adopting Adults' Stress Habits?, 2013. URL: https://www.apa.org/news/press/releases/stress/2013/stress-report.pdf. 174

[13] Viet Hoang Anh, Manh Ngo Van, Bang Ban Ha, and Thang Huynh Quyet. A real-time model based Support Vector Machine for emotion

recognition through EEG. In *2012 International Conference on Control, Automation and Information Sciences, ICCAIS 2012*, 2012. doi:10.1109/ICCAIS.2012.6466585. 158

[14] Emily Anthes. Pocket psychiatry: mobile mental-health apps have exploded onto the market, but few have been thoroughly tested. *Nature*, 532(7597):20–24, apr 2016. 13, 16

[15] Lisa Anthony, Sapna Prasad, Amy Hurst, and Ravi Kuber. A participatory design workshop on accessible apps and games with students with learning differences. In *ASSETS'12 - Proceedings of the 14th International ACM SIGACCESS Conference on Computers and Accessibility*, 2012. doi:10.1145/2384916.2384979. 55

[16] Brenna D. Argall and Aude G. Billard. A survey of Tactile HumanRobot Interactions, 2010. doi:10.1016/j.robot.2010.07.002. 44

[17] Ruben T Azevedo, Nell Bennett, Andreas Bilicki, Jack Hooper, Fotini Markopoulou, and Manos Tsakiris. The calming effect of a new wearable device during the anticipation of public speech. *Sci Rep*, 7(1):2285, 2017. 41, 167

[18] Muhammad Sufyian Mohd Azmi and Md Nasir Sulaiman. Accelerator-based human activity recognition using voting technique with NBTree and MLP classifiers. *International Journal on Advanced Science, Engineering and Information Technology*, 2017. doi:10.18517/ijaseit.7.1.1790. 92

[19] Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015. arXiv:1409.0473. 28

[20] Uddipana Baishya, Elgin Skye McLaren, and Alissa N. Antle. Idea bits: A tangible design tool to aid idea generation for tangible manipulation. In *Conference on Human Factors in Computing Systems - Proceedings*, 2019. doi:10.1145/3290607.3312820. 60

[21] Madeline Balaam, Geraldine Fitzpatrick, Judith Good, and Rosemary Luckin. Exploring Affective Technologies for the Classroom with the Subtle Stone. *Proceedings of the 28th international conference on Human factors in computing systems - CHI '10*, page 1623, 2009. 18

[22] Debrup Banerjee, Kazi Islam, Gang Mei, Lemin Xiao, Guangfan Zhang, Roger Xu, Shuiwang Ji, and Jiang Li. A deep transfer learning approach for improved post-traumatic stress disorder diagnosis. In *Proceedings - IEEE International Conference on Data Mining, ICDM*, 2017. 36

[23] Debrup Banerjee, Kazi Islam, Keyi Xue, Gang Mei, Lemin Xiao, Guangfan Zhang, Roger Xu, Cai Lei, Shuiwang Ji, and Jiang Li. A deep transfer learning approach for improved post-traumatic stress disorder diagnosis. *Knowledge and Information Systems*, 60(3):1693–1724, sep 2019. 36

[24] Ellen Banzhaf, Francisco De La Barrera, Annegret Kindler, Sonia Reyes-Paecke, Uwe Schlink, Juliane Welz, and Sigrun Kabisch. A conceptual framework for integrated analysis of environmental quality and quality of life. *Ecological Indicators*, 2014. doi:10.1016/j.ecolind.2014.06.002. 143

[25] Matthew Barker and LindenJanet. A sprite in the dark: supporting conventional mental healthcare practices with a tangible device. *Proceedings of the Tenth International Conference on Tangible, Embedded, and Embodied Interaction - TEI '17*, 2016. 18

[26] Paul D. Barrows and Shirley A. Thomas. Assessment of mood in aphasia following stroke: validation of the Dynamic Visual Analogue Mood Scales (D-VAMS). *Clinical Rehabilitation*, 32(1):94–102, jan 2018. doi:10.1177/0269215517714590. 111, 112

[27] A Batliner, S Steidl, and E Nöth. Releasing a thoroughly annotated and processed spontaneous emotional database: the FAU Aibo Emotion Corpus. *Proc. of Workshop on Corpora for Research on Emotion and Affect LREC*, 2008. 36

[28] David Beattie. SqueezeMusic- HCI Audio Interaction Research, 2017. 19

[29] Dror Ben-Zeev, Susan M. Kaiser, Christopher J. Brenner, Mark Begale, Jennifer Duffecy, and David C. Mohr. Development and usability testing of FOCUS: a smartphone system for self-management of schizophrenia. *Psychiatric rehabilitation journal*, 2013. 15

[30] Simone Benedetto, Christian Caldato, Elia Bazzan, Darren C. Greenwood, Virginia Pensabene, and Paolo Actis. Assessment of the fitbit charge 2 for monitoring heart rate. *PLoS ONE*, 2018. doi:10.1371/journal.pone.0192691. 4

[31] Ida H Bennike, Anders Wieghorst, and Ulrich Kirk. Online-based Mindfulness Training Reduces Behavioral Markers of Mind Wandering. 2017. 14

[32] Thomas Binder, Giorgio De Michelis, Pelle Ehn, Giulio Jacucci, Per Linde, and Ina Wagner. Participation in Design Things. In *Design Things*. 2019. doi:10.7551/mitpress/8262.003.0011. 53

[33] Martin Bollard and Paul Magee. Myhealth–developing accessible health materials with men with intellectual disability: a co-created Proof of Concept. *Design for Health*, 2020. doi:10.1080/24735132.2020.1836862. 59

[34] Fleur Heleen Boot, J. Owuor, J. Dinsmore, and M. Maclachlan. Access to assistive technology for people with intellectual disabilities: A systematic review to identify barriers and facilitators, 2018. doi:10.1111/jir.12532. 54

[35] Thomas Boraud. *How the Brain Makes Decisions*. 2020. doi:10.1093/oso/9780198824367.001.0001. 10

[36] Richard E. Boyatzis. Transforming qualitative information. *Sage Publications, Inc*, 1998. 64

[37] Jordan Boyd-Graber, Sonya Nikolova, Karyn Moffatt, Kenrick Kin, Joshua Lee, Lester Mackey, Marilyn Tremaine, and Maria Klawe. Participatory design with proxies: Developing a desktop-PDA system to support people with aphasia. In *Conference on Human Factors in Computing Systems - Proceedings*, 2006. 59

[38] Margaret M. Bradley and Peter J. Lang. Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 1994. doi:10.1016/0005-7916(94)90063-9. 11, 120

[39] José Bravo, Ramón Hervás, and Vladimir Villarreal. Ambient intelligence for health first international conference, AmIHEALTH 2015 Puerto Varas, Chile, December 1–4, 2015 proceedings. *Lecture Notes in Computer Science*, 9456:189–200, 2015. 46

[40] Margot Brereton, Laurianne Sitbon, Muhammad Haziq Lim Abdullah, Mark Vanderberg, and Stewart Koplick. Design after design to bridge between people living with cognitive or sensory impairments, their friends and proxies. *CoDesign*, 2015. doi:10.1080/15710882.2015.1009471. 58

[41] Amanda Briggs-Goode, Martha Glazzard, Sarah Walker, Sarah Kettley, Tincuta Heinzel, and Rachel Lucas. Wellbeing and Smart Textiles: Reflecting on Collaborative Practices and the Design Process. *Journal of Textile Design Research and Practice*, 2016. doi:10.1080/20511787.2016.1256599. 60

[42] D Brown, R Kettley, and S Kettley. Interpersonal Process Recall – a method for reflecting on co-design experiences. In *Method Stories Workshop, ACM CHI 2016*, 2016. URL: https://chimethodstories.wordpress.com/. 57

[43] D Brown, R Kettley, and S Kettley. Interpersonal Process Recall – a method for reflecting on co-design experiences. 2016. 86

[44] David J. Brown, Steven Battersby, Penny Standen, Nicola Anderton, Matthew Harrison, and David J. Brown. Inclusive design for disability: A prototype switch for use in virtual learning environments. *International Journal on Disability and Human Development*, 2005. doi:10.1515/IJDHD.2005.4.2.103. 60, 74

[45] David J. Brown, David McHugh, Penny Standen, Lindsay Evett, Nick Shopland, and Steven Battersby. Designing location-based learning experiences for people with intellectual disabilities and additional sensory impairments.

Computers and Education, 2011. doi:10.1016/j.compedu.2010.04.014. 57

[46] R. M. Bruno. Interpretation of pictorially presented social situations by learning disabled and normal children. *Journal of learning disabilities*, 1981. doi:10.1177/002221948101400617. 53

[47] Miguel Bruns, Alonso David, V Keyson, and Caroline C M Hummels. Squeeze, Rock, and Roll; Can Tangible Interaction with Affective Products Support Stress Reduction? *Proceedings of the 2nd international conference on Tangible and embedded interaction - TEI '08*, 2008. 40

[48] Philip M. Burgess, Jane E. Pirkis, Tim N. Slade, Amy K. Johnston, Graham N. Meadows, and Jane M. Gunn. Service Use for Mental Health Problems: Findings from the 2007 National Survey of Mental Health and Wellbeing. *Australian New Zealand Journal of Psychiatry*, 43(7):615–623, jul 2009. doi:10.1080/00048670902970858. 13

[49] Ingrid Burkett. An Introduction to Co-design. Technical report, 2016. URL: http://www.csi.edu.au/. 53, 58

[50] Steven H Cady and Gwen E Jones. Massage therapy as a workplace intervention for reduction of stress. *O Perceptual and Motor Skills*, 84:157–158, 1997. 41

[51] Rafael A. Calvo and Dorian Peters. Promoting psychological wellbeing: Loftier goals for new technologies [opinion], 2013. doi:10.1109/MTS.2013.2286429. 52

[52] Yekta Said Can, Bert Arnrich, and Cem Ersoy. Stress detection in daily life scenarios using smart phones and wearable sensors: A survey. *Journal of Biomedical Informatics*, 2019. doi:10.1016/j.jbi.2019.103139. 38, 87

[53] Vaughan Carr. Patients' techniques for coping with schizophrenia: An exploratory study. *British Journal of Medical Psychology*, 61(4):339–352, dec 1988. 40, 161

[54] Rich Caruana. Multitask Learning. *Machine Learning*, 28(1):41–75, 1997. doi:10.1023/A:1007379606734. 35, 118, 141

[55] R. Castaldo, W. Xu, P. Melillo, L. Pecchia, L. Santamaria, and C. James. Detection of mental stress due to oral academic examination via ultra-short-term HRV analysis. In *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, 2016. doi:10.1109/EMBC.2016.7591557. 37, 157

[56] Nair Chaitanya M., Sharmila Jayakkumar, Edwin Chong, and C.H. Yeow. A wearable, EEG-based massage headband for anxiety alleviation. In *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, volume 2017, pages 3557–3560. IEEE, jul 2017. doi:10.1109/EMBC.2017.8037625. 46

[57] Hsiu-Sen Chiang. ECG-based Mental Stress Assessment Using Fuzzy Computing and Associative Petri Net. *Journal of Medical and Biological Engineering*, 35(6):833–844, dec 2015. doi:10.1007/s40846-015-0095-7. 20

[58] Jongyoon Choi, B. Ahmed, and Ricardo Gutierrez-Osuna. Development and evaluation of an ambulatory stress monitor based on wearable sensors. *IEEE Transactions on Information Technology in Biomedicine*, 2012. doi:10.1109/TITB.2011.2169804. 20

[59] Yoorim Choi and Minjung Kim. Measurement of occupants' stress based on electroencephalograms (EEG) in twelve combined environments. *Building and Environment*, 88:65–72, jun 2015. doi:10.1016/J.BUILDENV.2014.10.003. 33

[60] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. dec 2014. URL: http://arxiv.org/abs/1412.3555, arXiv:1412.3555. 26

[61] S. Clement, O. Schauman, T. Graham, F. Maggioni, S. Evans-Lacko, N. Bezborodovs, C. Morgan, N. Rüsch, J. S. L. Brown, and G. Thornicroft. What is the impact of mental health-related stigma on help-seeking? A systematic review of quantitative and qualitative studies. *Psychological Medicine*, 45(01):11–27, jan 2015. 13, 51

[62] Brendan Corbett, Chang S. Nam, and Takehiko Yamaguchi. The Effects of Haptic Feedback and Visual Distraction on Pointing Task Performance. *International Journal of Human-Computer Interaction*, 32(2):89–102, feb 2016. 40, 167

[63] Jean Costa, François Guimbretière, Malte Jung, and Tanzeem Choudhury. BoostMeUp: Improving Cognitive Performance in the Moment by Unobtrusively Regulating Emotions with a Smartwatch. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2019. doi:10.1145/3328911. 165

[64] Ian Cross, Susan Hallam, and Michael Thaut. *The Oxford Handbook of Music Psychology*. 2008. doi:10.1093/oxfordhb/9780199298457.001.0001. 165

[65] Suzanne B. Da Câmara, Rakshit Agrawal, and Katherine Isbister. Identifying children's fidget object preferences: Toward exploring the impacts of fidgeting and fidget-friendly tangibles. In *DIS 2018 - Proceedings of the 2018 Designing Interactive Systems Conference*, 2018. doi:10.1145/3196709.3196790. 80, 161

[66] Simon D'Alfonso, Olga Santesteban-Echarri, Simon Rice, Greg Wadley, Reeva Lederman, Christopher Miles, John Gleeson, and Mario Alvarez-Jimenez. Artificial Intelligence-Assisted Online Social Therapy for Youth Mental Health. *Frontiers in Psychology*, 8:796, jun 2017. 14

[67] Antonio R. Damasio. Emotion in the perspective of an integrated nervous system. In *Brain Research Reviews*, volume 26, 1998. doi:10.1016/S0165-0173(97)00064-7. 10

[68] Priyanka Das, Anwesha Khasnobish, and D. N. Tibarewala. Emotion recognition employing ECG and GSR signals as markers of ANS. In *Conference on Advances in Signal Processing, CASP 2016*, 2016. doi:10.1109/CASP.2016.7746134. 32

[69] Claudia Daudén Roquet and Corina Sas. Evaluating Mindfulness Meditation Apps. In *Extended Abstracts of the 2018 CHI Conference on Human*

Factors in Computing Systems - CHI '18, pages 1–6, New York, New York, USA, 2018. ACM Press. URL: http://dl.acm.org/citation.cfm?doid=3170427.3188616, doi:10.1145/3170427.3188616. 161

[70] Dmitry Davidov, Oren Tsur, and Ari Rappoport. Enhanced Sentiment Learning Using Twitter Hashtags and Smileys. Technical report, 2010. URL: https://www.aclweb.org/anthology/C10-2028. 21

[71] Melissa Dawe. "let me show you what i want": Engaging individuals with cognitive disabilities and their families in design. In *Conference on Human Factors in Computing Systems - Proceedings*, 2007. doi:10.1145/1240866.1240976. 55

[72] Thomas De Cooman, Kaat Vandecasteele, Carolina Varon, Borbála Hunyadi, Evy Cleeren, Wim Van Paesschen, and Sabine Van Huffel. Personalizing Heart Rate-Based Seizure Detection Using Supervised SVM Transfer Learning. *Frontiers in Neurology*, 2020. doi:10.3389/fneur.2020.00145. 35

[73] Alberto De Santos Sierra, Carmen Sanchez Avila, Javier Guerra Casanova, Gonzalo Bailador Del Pozo, and Vicente Jara Vera. Two stress detection schemes based on physiological signals for real-time applications. In *Proceedings - 2010 6th International Conference on Intelligent Information Hiding and Multimedia Signal Processing, IIHMSP 2010*, 2010. doi:10.1109/IIHMSP.2010.95. 33

[74] Katarina Dedovic, Robert Renwick, Najmeh Khalili Mahani, and Veronika Engert. The Montreal Imaging Stress Task : using functional imaging to investigate the ... *Psychiatry Neuroscience*, 30(5):319–325, 2005. 120

[75] Janez Demšar. Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal of Machine Learning Research*, 7(Jan):1–30, 2006. URL: http://www.jmlr.org/papers/v7/demsar06a.html. 103

[76] Jun Deng, Sascha Fruhholz, Zixing Zhang, and Bjorn Schuller. Recognizing Emotions from Whispered Speech Based on Acoustic Feature Transfer Learning. *IEEE Access*, pages 1–1, 2017. doi:10.1109/ACCESS.2017.2672722. 36

[77] Jun Deng, Zixing Zhang, Erik Marchi, and Björn Schuller. Sparse autoencoder-based feature transfer learning for speech emotion recognition. In *Proceedings - 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, ACII 2013*, pages 511–516, 2013. doi:10.1109/ACII.2013.90. 36

[78] David DeSteno, Daniel Lim, Fred Duong, and Paul Condon. Meditation Inhibits Aggressive Responses to Provocations. *Mindfulness*, 9(4):1117–1122, aug 2018. 14

[79] Gavin Doherty, David Coyle, John Sharry, and James Joyce St. *Engagement with Online Mental Health Interventions: An Exploratory Clinical Study of a Treatment for Depression*. 2012. URL: https://www.scss.tcd.ie/Gavin.Doherty/papers/DohertyCS12.pdf. 161

[80] Tara Donker, Katherine Petrie, Judy Proudfoot, Janine Clarke, Mary-Rose Birch, and Helen Christensen. Smartphones for smarter delivery of mental health programs: a systematic review. *Journal of medical Internet research*, 15(11):e247, nov 2013. doi:10.2196/jmir.2791. 16

[81] Marcos Economides, Janis Martman, Megan J. Bell, and Brad Sanderson. Improvements in Stress, Affect, and Irritability Following Brief Use of a Mindfulness-based Smartphone App: A Randomized Controlled Trial. *Mindfulness*, 9(5):1584–1593, oct 2018. 14

[82] Bo Edvardsson, Thomas Meiren, Adrienne Schäfer, and Lars Witell. Having a strategy for new service development - does it really matter? *Journal of Service Management*, 2013. doi:10.1108/09564231311304170. 54

[83] Paul Ekman and Wallace V. Friesen. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 1971. doi:10.1037/h0030377. 11

[84] Nour El Mawass and Eiman Kanjo. A supermarket stress map. In *UbiComp 2013 Adjunct - Adjunct Publication of the 2013 ACM Conference on Ubiquitous Computing*, 2013. 2

[85] Irene Elkin, M. Tracie Shea, John T. Watkins, Stanley D. Imber, Stuart M. Sotsky, Joseph F. Collins, David R. Glass, Paul A. Pilkonis, William R. Leber, John P. Docherty, Susan J. Fiester, and Morris B. Parloff. National Institute of Mental Health Treatment of Depression Collaborative Research Program. *Archives of General Psychiatry*, 46(11):971, nov 1989. doi:10.1001/archpsyc.1989.01810110013002. 13

[86] Empatica. E4 wristband | Real-time physiological signals | Wearable PPG, EDA, Temperature, Motion sensors, 2020. 49

[87] European Comission. Data protection, 2018. URL: https://ec.europa.eu/info/law/law-topic/data-protection_en. 46

[88] M Everingham, L Van Gool, C K I Williams, J Winn, and A Zisserman. DL PASCAL The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Development Kit. *Pattern Analysis, Statistical Modelling and Computational Learning, Tech. Rep*, 2012. 35

[89] Fatemeh Fahimi, Zhuo Zhang, Wooi Boon Goh, Tih Shi Lee, Kai Keng Ang, and Cuntai Guan. Inter-subject transfer learning with an end-to-end deep convolutional neural network for EEG-based BCI. *Journal of Neural Engineering*, 2019. doi:10.1088/1741-2552/aaf3f6. 35

[90] James Farley, Evan F. Risko, and Alan Kingstone. Everyday attention and lecture retention: The effects of time, fidgeting, and mind wandering. *Frontiers in Psychology*, 2013. doi:10.3389/fpsyg.2013.00619. 161

[91] Hassan Ismail Fawaz, Benjamin Lucas, Germain Forestier, Charlotte Pelletier, Daniel F. Schmidt, Jonathan Weber, Geoffrey I. Webb, Lhassane Idoumghar, Pierre-Alain Muller, and François Petitjean. InceptionTime: Finding AlexNet for Time Series Classification. sep 2019. arXiv:1909.04939. xii, 29

[92] Claire L Forrest, Jenny L Gibson, Sarah L Halligan, and Michelle C St Clair. A longitudinal analysis of early language difficulty and peer problems on later emotional difficulties in adolescence: Evidence from the Millennium Cohort Study. *Autism Developmental Language Impairments*, 2018. 166

[93] Peter Francis, Sandrine Balbo, and Lucy Firth. Towards co-design with users who have autism spectrum disorders. *Universal Access in the Information Society*, 2009. doi:10.1007/s10209-008-0143-y. 54

[94] Jérémy Frey, Jessica Cauchard, and I D C Herzliya. Remote Biofeedback Sharing, Opportunities and Challenges. *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 730–733, aug 2018. arXiv:1808.08713, doi:arXiv:1808.08713v1. 41, 45, 46

[95] Carolina Fuentes, Iyubanit Rodríguez, and Valeria Herskovic. EmoBall: A study on a tangible interface to self-report emotional information considering digital competences. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2015. doi:10.1007/978-3-319-26508-7_19. 17, 47, 57

[96] T. L. Galinsky, R. R. Rosa, J. S. Warm, and W. N. Dember. Psychophysical determinants of stress in sustained attention. *Human Factors*, 1993. doi:10.1177/001872089303500402. 161

[97] Julia Galliers, Stephanie Wilson, Abi Roper, Naomi Cocks, Jane Marshall, Sam Muscroft, and Tim Pring. Words are not enough: Empowering people with aphasia in the design process. In *ACM International Conference Proceeding Series*, 2012. doi:10.1145/2347635.2347643. 59

[98] 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, dec 2018. 4, 118

[99] B. Geethanjali, K. Adalarasu, A. Hemapraba, S. Pravin Kumar, and R. Rajasekeran. Emotion analysis using sam (Self-assessment manikin) scale. *Biomedical Research (India)*, 2017. 112

[100] Peter J. Gianaros, J. Richard Jennings, Lei K. Sheu, Phil J. Greer, Lewis H. Kuller, and Karen A. Matthews. Prospective reports of chronic life stress predict decreased grey matter volume in the hippocampus. *NeuroImage*, 2007. doi:10.1016/j.neuroimage.2006.10.045. 39

[101] Giorgos Giannakakis, Kostas Marias, and Manolis Tsiknakis. A stress recognition system using HRV parameters and machine learning techniques. In *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos, ACIIW 2019*, pages 269–272. Institute of Electrical and Electronics Engineers Inc., sep 2019. doi:10.1109/ACIIW.2019.8925142. 32

[102] Alison Giraud-Saunders. Mental health in people with learning disabilities, 2011. 1, 6, 51, 52, 82, 118, 172

[103] Martin Gjoreski, Hristijan Gjoreski, Mitja Luštrek, and Matjaž Gams. Continuous stress detection using a wrist device - in laboratory and real life. In *UbiComp 2016 Adjunct - Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2016. doi:10.1145/2968219.2968306. 37

[104] Martin Gjoreski, Mitja Luštrek, Matjaž Gams, and Hristijan Gjoreski. Monitoring stress with a wrist device using context. *Journal of Biomedical Informatics*, 2017. doi:10.1016/j.jbi.2017.08.006. 12

[105] David S. Goldstein and Irwin J. Kopin. Evolution of concepts of stress, 2007. doi:10.1080/10253890701288935. 12

[106] HARRIS S. GOLDSTEIN and ROBERT EDELBERG. A plethysmographic method for demonstrating the response specificity of the oral vascular bed. *Psychophysiology*, 34(1):124–128, jan 1997. doi:10.1111/j.1469-8986.1997.tb02423.x. 20

[107] Ian Goodfellow, Yoshua Benigo, and Courville Aaron. Deep Learning (Adaptive Computation and Machine Learning), 2016. 118

[108] Google. Google Play Store, 2019. URL: https://play.google.com/store/apps?hl=en. 88

[109] Google. reCAPTCHA: Easy on Humans, Hard on Bots, 2019. URL: https://www.google.com/recaptcha/intro/v3.html. 21

[110] Jack M. Gorman and Richard P. Sloan. Heart rate variability in depressive and anxiety disorders. *American Heart Journal*, 2000. doi:10.1067/mhj.2000.109981. 20

[111] D. Gould. Visual Analogue Scale (VAS) - Information Point. *Journal of Clinical Nursing*, 2001. doi:10.4193/Rhino10.303. 120

[112] Alessandra Grassi, Andrea Gaggioli, and Giuseppe Riva. New technologies to manage exam anxiety. *Annual Review of CyberTherapy and Telemedicine*, 2011. 15

[113] Matthew M Graziose. On the Accuracy of Self-Report Instruments for Measuring Food Consumption in the School Setting. *Advances in Nutrition*, 8(4):635–636, jul 2017. 12, 45

[114] James J Gross and R Thompson. Emotion regulation: Conceptual foundations. In *J. J. Gross (Eds.) Handbook of Emotion Regulation*. 2007. 52

[115] Frode Guribye and Tor Gjøsæter. Tangible Interaction in the Dentist Office. In *Proceedings of the Twelfth International Conference on Tangible, Embedded, and Embodied Interaction - TEI '18*, pages 123–130, New York, New York, USA, 2018. ACM Press. doi:10.1145/3173225.3173287. 19

[116] John Garcia Gustavson and Andrew R. The Science of Self-Report. *APS Observer*, 10(1), jan 1997. 12

[117] Morgane Hamon, Rémy Ramadour, and Jérémy Frey. Exploring biofeedback with a tangible interface designed for relaxation. In *PhyCS 2018 - Proceedings of the 5th International Conference on Physiological Computing Systems*, 2018. arXiv:1808.08711, doi:10.5220/0006961200540063. 42

[118] Gordon T. Harold and Dale F. Hay. Normal development in middle childhood. *Psychiatry*, 4(6):3–5, jun 2005. doi:10.1383/psyt.4.6.3.66355. 77

[119] Virginia Harrison, Judith Proudfoot, Pang Ping Wee, Gordon Parker, Dusan Hadzi Pavlovic, and Vijaya Manicavasagar. Mobile mental health: Review of the emerging field and proof of concept study. *Journal of Mental*

Health, 20(6):509–524, dec 2011. doi:10.3109/09638237.2011.608746. 17

[120] Sandra G. Hart and Lowell E. Staveland. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. *Advances in Psychology*, 1988. doi:10.1016/S0166-4115(08)62386-9. 120

[121] Md Junayed Hasan, Muhammad Sohaib, and Jong Myon Kim. 1D CNN-based transfer learning model for bearing fault diagnosis under variable working conditions. In *Advances in Intelligent Systems and Computing*, volume 888, pages 13–23. Springer Verlag, nov 2019. doi:10.1007/978-3-030-03302-6_2. 23

[122] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision*, 2015. arXiv:1502.01852, doi:10.1109/ICCV.2015.123. 28

[123] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016. arXiv:1512.03385, doi:10.1109/CVPR.2016.90. 28

[124] J.A. Healey and R.W. Picard. Detecting Stress During Real-World Driving Tasks Using Physiological Sensors. *IEEE Transactions on Intelligent Transportation Systems*, 6(2):156–166, jun 2005. 32, 34, 46

[125] Jennifer Healey, Lama Nachman, Sushmita Subramanian, Junaith Shahabdeen, and Margaret Morris. Out of the lab and into the fray: Towards modeling emotion in everyday life. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2010. doi:10.1007/978-3-642-12654-3_10. 38

[126] Niels Hendriks, Karin Slegers, and Pieter Duysburgh. Codesign with people living with cognitive or sensory impairments: a case for method stories and uniqueness. *International Journal of CoCreation in Design and the Arts*, 2015. 58

[127] Katherine A. Herborn, James L. Graves, Paul Jerem, Neil P. Evans, Ruedi Nager, Dominic J. McCafferty, and Dorothy E.F. F McKeegan. Skin temperature reveals the intensity of acute stress. *Physiology Behavior*, 152(Pt A):225–230, dec 2015. doi:10.1016/J.PHYSBEH.2015.09.032. 33

[128] David Hernando, Surya Roca, Jorge Sancho, Álvaro Alesanco, and Raquel Bailón. Validation of the apple watch for heart rate variability measurements during relax and mental stress in healthy subjects. *Sensors (Switzerland)*, 2018. doi:10.3390/s18082619. 4

[129] Geoffrey E. Hinton, Alex Krizhevsky, and Sida D. Wang. Transforming auto-encoders. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2011. 26

[130] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, nov 1997. doi:10.1162/neco.1997.9.8.1735. 24, 98

[131] Arthur E. Hoerl and Robert W. Kennard. Ridge Regression: Applications to Nonorthogonal Problems. *Technometrics*, 1970. doi:10.1080/00401706.1970.10488635. 28

[132] M A Hogg, Dominic Abrams, and G N Martin. Social cognition and attitudes. *Psychology*, pages 646–677, 2010. 10

[133] Nic Hollinworth, Gosia Malgosia Kwiatkowska, Faustina Hwang, Andy Minnion, and Kate Allen. Making electronics more accessible to people with learning disabilities. In *Conference on Human Factors in Computing Systems - Proceedings*, 2014. doi:10.1145/2559206.2581175. 60

[134] Stefan Holmlid. Participative, co-operative, emancipatory: From participatory design to service design. *First Nordic Conference on Service Design and Service Innovation*, 2009. 53

[135] Kristina Höök, Martin P. Jonsson, Anna Ståhl, and Johanna Mercurio. Somaesthetic Appreciation design. In *Conference on Human Factors in*

Computing Systems - Proceedings, 2016. doi:10.1145/2858036.2858583. 42

[136] Eva Hornecker and Jacob Buur. Getting a Grip on Tangible Interaction: A Framework on Physical Space and Social Interaction. *Proceedings of the SIGCHI conference on Human Factors in computing systems - CHI '06*, 2006. doi:10.1145/????????? 52

[137] Rasam Bin Hossain, Mefta Sadat, and Hasan Mahmud. Recognition of human affection in smartphone perspective based on accelerometer and user's sitting position. In *2014 17th International Conference on Computer and Information Technology, ICCIT 2014*, 2014. 21

[138] Karen Hovsepian, Mustafa Al'absi, Emre Ertin, Thomas Kamarck, Motohiro Nakajima, and Santosh Kumar. CStress: Towards a gold standard for continuous stress assessment in the mobile environment. In *UbiComp 2015 - Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2015. doi:10.1145/2750858.2807526. 37, 157

[139] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V Le, Hartwig Adam, Google Ai, and Google Brain. Searching for MobileNetV3. Technical report. arXiv:1905.02244v5. 35

[140] Guang Bin Huang, Hongming Zhou, Xiaojian Ding, and Rui Zhang. Extreme learning machine for regression and multiclass classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 2012. doi:10.1109/TSMCB.2011.2168604. 36

[141] B. F. Hudson, J. Ogden, and M. S. Whiteley. Randomized controlled trial to compare the effect of simple distraction interventions on pain and anxiety experienced during conscious surgery. *European Journal of Pain (United Kingdom)*, 2015. 2, 80, 138, 161

[142] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *32nd International Conference on Machine Learning, ICML 2015*, volume 1, pages 448–456. International Machine Learning Society (IMLS), 2015. 24, 98

[143] Hiroshi Ishii. The tangible user interface and its evolution. *Communications of the ACM*, 2008. doi:10.1145/1349026.1349034. 2

[144] Hiroshi Ishii and Brygg Ullmer. Tangible bits: Towards seamless interfaces between people, bits and atoms. In *Conference on Human Factors in Computing Systems - Proceedings*, 1997. 2

[145] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre Alain Muller. Deep learning for time series classification: a review. *Data Mining and Knowledge Discovery*, 2019. arXiv:1809.04356, doi:10.1007/s10618-019-00619-1. 127

[146] Robert J.K. Jacob, Orit Shaer, Audrey Girouard, Leanne M. Hirshfield, Michael S. Horn, Erin Treacy Solovey, and Jamie Zigelbaum. Reality-Based interaction: A framework for post-WIMP interfaces. In *Conference on Human Factors in Computing Systems - Proceedings*, 2008. doi:10.1145/1357054.1357089. 17

[147] Natasha Jaques, Sara Taylor, Ehimwenma Nosakhare, Akane Sano, and Rosalind Picard. Multi-task Learning for Predicting Health, Stress, and Happiness, 2016. 38

[148] Noppadon Jatupaiboon, Setha Pan-Ngum, and Pasin Israsena. Real-time EEG-based happiness detection system. *The Scientific World Journal*, 2013. doi:10.1155/2013/618649. 158

[149] Houtan Jebelli, Sungjoo Hwang, and Sang Hyun Lee. EEG-based workers' stress recognition at construction sites. *Automation in Construction*, 2018. doi:10.1016/j.autcon.2018.05.027. 33

[150] Sooyeon Jeong, Miriam Zisook, Luke Plummer, Cynthia Breazeal, Peter Weinstock, Deirdre E. Logan, Matthew S. Goodwin, Suzanne Graca, Brianna O'Connell, Honey Goodenough, Laurel Anderson, Nicole Stenquist, and Katie Fitzpatrick. A Social Robot to Mitigate Stress, Anxiety, and Pain in Hospital Pediatric Care. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts - HRI'15 Extended Abstracts*, pages 103–104, New York, New York, USA, 2015. ACM Press. doi:10.1145/2701973.2702028. 44

[151] Jutta Joormann, Matthias Siemer, and Ian H Gotlib. Mood Regulation in Depression: Differential Effects of Distraction and Recall of Happy Memories on Sad Mood. 2007. 161

[152] Rafal Jozefowicz, Wojciech Zaremba, and Ilya Sutskever. An empirical exploration of recurrent network architectures, 2015. URL: https://dl.acm.org/citation.cfm?id=3045367. 26

[153] Łukasz Kaiser and Ilya Sutskever. Neural GPUs Learn Algorithms. nov 2015. URL: http://arxiv.org/abs/1511.08228, arXiv:1511.08228. 26

[154] Eiman Kanjo, Daria J. Kuss, and Chee Siang Ang. NotiMind: Utilizing Responses to Smart Phone Notifications as Affective Sensors. *IEEE Access*, 5:22023–22035, 2017. doi:10.1109/ACCESS.2017.2755661. 13, 21

[155] Eiman Kanjo, Eman M G Younis, and Chee Siang Ang. Deep learning analysis of mobile physiological, environmental and location sensor data for emotion detection. *Information Fusion*, 49:46–56, 2019. 34, 157, 159

[156] Eiman Kanjo, Eman M.G. Younis, and Nasser Sherkat. Towards unravelling the relationship between on-body, environmental and emotion data using sensor information fusion approach. *Information Fusion*, 2018. 21, 34, 157, 159

[157] Saeed Karimi-Bidhendi, Faramarz Munshi, and Ashfaq Munshi. Scalable Classification of Univariate and Multivariate Time Series. In *Proceedings - 2018 IEEE International Conference on Big Data, Big Data 2018*, 2019. doi:10.1109/BigData.2018.8621889. 141, 147

[158] Connor Tom Keating and Jennifer Louise Cook. Facial Expression Production and Recognition in Autism Spectrum Disorders: A Shifting Landscape, 2020. doi:10.1016/j.chc.2020.02.006. 112

[159] Chelsea Kelling, Daniella Pitaro, and Jussi Rantala. Good vibes. In *Proceedings of the 20th International Academic Mindtrek Conference on - AcademicMindtrek '16*, pages 130–136, New York, New York, USA, 2016. ACM Press. 40, 165

[160] A. M. Khan, Y. K. Lee, and S. Y. Lee. Accelerometer's position free human activity recognition using a hierarchical recognition model. In *12th IEEE International Conference on e-Health Networking, Application and Services, Healthcom 2010*, 2010. doi:10.1109/HEALTH.2010.5556553. 92

[161] Angela S. Khor, Kylie M. Gray, Sophie C. Reid, and Glenn A. Melvin. Feasibility and validity of ecological momentary assessment in adolescents with high-functioning autism and Asperger's disorder. *Journal of Adolescence*, 2014. doi:10.1016/j.adolescence.2013.10.005. 107

[162] Reza Khosrowabadi, Chai Quek, Kai Keng Ang, Sau Wai Tung, and Michel Heijnen. A Brain-Computer Interface for classifying EEG correlates of chronic mental stress. In *The 2011 International Joint Conference on Neural Networks*, pages 757–762. IEEE, jul 2011. doi:10.1109/IJCNN.2011.6033297. 30, 33, 34, 46

[163] George Poonkhin Khut. Designing biofeedback artworks for relaxation. In *Conference on Human Factors in Computing Systems - Proceedings*, 2016. doi:10.1145/2851581.2891089. 42

[164] Basel Kikhia, Thanos G Stavropoulos, Stelios Andreadis, Niklas Karvonen, Ioannis Kompatsiaris, Stefan Sävenstedt, Marten Pijl, and Catharina Melander. Utilizing a Wristband Sensor to Measure the Stress Level for People with Dementia. *Sensors (Basel, Switzerland)*, 16(12), nov 2016. doi:10.3390/s16121989. 33

[165] Hye-Geum Kim, Eun-Jin Cheon, Dai-Seg Bai, Young Hwan Lee, and Bon-Hoon Koo. Stress and Heart Rate Variability: A Meta-Analysis and Review of the Literature. *Psychiatry investigation*, 15(3):235–245, mar 2018. 2

[166] Jonghwa Kim and Elisabeth André. Emotion recognition based on physiological changes in music listening. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2008. doi:10.1109/TPAMI.2008.26. 37

[167] Nobuaki Kimura, Ikuo Yoshinaga, Kenji Sekijima, Issaku Azechi, and Daichi Baba. Convolutional Neural Network Coupled with a Transfer-Learning Approach for Time-Series Flood Predictions. *Water*, 12(1):96, dec 2019. 125

[168] Evan Kleiman, Alexander J. Millner, Victoria W. Joyce, Carol C. Nash, Ralph J. Buonopane, and Matthew K. Nock. Using wearable physiological monitors with suicidal adolescent inpatients: Feasibility and acceptability study. *JMIR mHealth and uHealth*, 2019. doi:10.2196/13725. 4

[169] Saskia Koldijk, Maya Sappelli, Suzan Verberne, Mark A. Neerincx, and Wessel Kraaij. The Swell knowledge work dataset for stress and user modeling research. In *ICMI 2014 - Proceedings of the 2014 International Conference on Multimodal Interaction*, pages 291–298. Association for Computing Machinery, Inc, nov 2014. doi:10.1145/2663204.2663257. 3, 108, 122

[170] Yuka Kotozaki, Hikaru Takeuchi, Atsushi Sekiguchi, Yuki Yamamoto, Takamitsu Shinada, Tsuyoshi Araki, Kei Takahashi, Yasuyuki Taki, Takeshi Ogino, Masashi Kiguchi, and Ryuta Kawashima. Biofeedback-based training for stress management in daily hassles: an intervention study. *Brain and behavior*, 4(4):566–579, jul 2014. doi:10.1002/brb3.241. 39

[171] 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, nov 2009. 31

[172] Jacob Leon Kröger, Otto Hans Martin Lutz, and Philip Raschke. Privacy implications of voice and speech analysis – information disclosure by inference. In *IFIP Advances in Information and Communication Technology*, 2020. doi:10.1007/978-3-030-42504-3_16. 77

[173] Dustin C. Krutsinger, Kuldeep N. Yadav, Elizabeth Cooney, Steven Brooks, Scott D. Halpern, and Katherine R. Courtright. A pilot randomized trial of five financial incentive strategies to increase study enrollment and retention rates. *Contemporary Clinical Trials Communications*, 2019. doi:10.1016/j.conctc.2019.100390. 141

[174] Naoko Kudo, Shinohara Hitomi, and Hideya Kodama. Heart Rate Variability Biofeedback Intervention for Reduction of Psychological Stress During the Early Postpartum Period. *Applied Psychophysiology and Biofeedback*, 39(3-4):203–211, 2014. 39

[175] Vivian Kwan, Gregory Hagen, Melanie Noel, Keith Dobson, and Keith Yeates. Healthcare at Your Fingertips: The Professional Ethics of Smartphone Health-Monitoring Applications. *Ethics Behavior*, 27(8):615–631, nov 2017. URL: https://www.tandfonline.com/doi/full/10.1080/10508422.2017.1285237, doi:10.1080/10508422.2017.1285237. 21

[176] Jennifer R. Kwapisz, Gary M. Weiss, and Samuel A. Moore. Activity recognition using cell phone accelerometers. *ACM SIGKDD Explorations Newsletter*, 2011. doi:10.1145/1964897.1964918. 22, 141, 142, 146, 173

[177] Lan lan Chen, Ao Zhang, and Xiao guang Lou. Cross-subject driver status detection from physiological signals based on hybrid feature selection and transfer learning. *Expert Systems with Applications*, 2019. doi:10.1016/j.eswa.2019.02.005. 35

[178] Óscar D. Lara and Miguel A. Labrador. A survey on human activity recognition using wearable sensors. *IEEE Communications Surveys and Tutorials*, 15(3):1192–1209, 2013. doi:10.1109/SURV.2012.110112.00192. 22, 88

[179] Walter S. Lasecki, Young Chol Song, Henry Kautz, and Jeffrey P. Bigham. Real-time crowd labeling for deployable activity recognition. In *Proceedings of the 2013 conference on Computer supported cooperative work - CSCW '13*, page 1203, New York, New York, USA, 2013. ACM Press. doi:10.1145/2441776.2441912. 21, 88

[180] Siddique Latif, Rajib Rana, Shahzad Younis, Junaid Qadir, and Julien Epps. Transfer learning for improving speech emotion classification accuracy. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2018. arXiv:1801.06353, doi:10.21437/Interspeech.2018-1625. 36

[181] Amanda Lazar, Jessica L. Feuston, Caroline Edasis, and Anne Marie Piper. Making as expression: Informing design with people with complex communication needs through art therapy. In *Conference on Human Factors in Computing Systems - Proceedings*, 2018. doi:10.1145/3173574.3173925. 55

[182] James Alexander Lee, Lu Bai, and Christos Efstratiou. OSN mood tracking: Exploring the use of online social network activity as an indicator of mood changes. In *UbiComp 2016 Adjunct - Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2016. doi:10.1145/2968219.2968304. 30

[183] James Alexander Lee, Christos Efstratiou, Panote Siriaraya, Dinkar Sharma, and Chee Siang Ang. SnapAppy: A positive psychology intervention using smartphone photography to improve emotional well-being. *Pervasive and Mobile Computing*, page 101369, mar 2021. URL: https://linkinghub.elsevier.com/retrieve/pii/S1574119221000377, doi:10.1016/j.pmcj.2021.101369. 42

[184] Hong Li, Jonna Häkkilä, and Kaisa Väänänen. Review of unconventional user interfaces for emotional communication between long-distance partners. In *MobileHCI 2018 - Beyond Mobile: The Next 20 Years - 20th International Conference on Human-Computer Interaction with Mobile Devices and Services, Conference Proceedings*, 2018. doi:10.1145/3229434.3229467. 41

[185] Jinpeng Li, Shuang Qiu, Yuan-Yuan Shen, Cheng-Lin Liu, and Huiguang He. Multisource Transfer Learning for Cross-Subject EEG Emotion Recognition. *IEEE Transactions on Cybernetics*, 50(7):3281 – 3293, 2019. 36, 117

[186] Rong-Hao Liang, Bin Yu, Mengru Xue, Jun Hu, and Loe M G Feijs. BioFidget: Biofeedback for Respiration Training Using an Augmented Fidget Spinner. *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, 2018. doi:10.1145/3173574.3174187. 42, 46

[187] Berta Liliana and Brusilovsky Filer. *Evaluating Cognitive Accessibility. Scientific keys to strengthen the role of the evaluator with functional diversity (English edition).* La Ciudad Accesible, 2018. 82

[188] Daniel Lim, Paul Condon, and David DeSteno. Mindfulness and Compassion: An Examination of Mechanism and Scalability. *PLOS ONE*, 10(2):e0118221, feb 2015. 14

[189] Yuan Pin Lin and Tzyy Ping Jung. Improving EEG-based emotion classification using conditional transfer learning. *Frontiers in Human Neuroscience*, 2017. 36

[190] Yuan Pin Lin, Chi Hong Wang, Tzyy Ping Jung, Tien Lin Wu, Shyh Kang Jeng, Jeng Ren Duann, and Jyh Horng Chen. EEG-based emotion recognition in music listening. *IEEE Transactions on Biomedical Engineering*, 2010. doi:10.1109/TBME.2010.2048568. 158

[191] Geoff Lindsay and Julie Dockrell. The relationship between speech, language and communication needs (SLCN) and behavioural, emotional and social difficulties (BESD). *Department for Education*, 2012. 166

[192] Christine Lætitia Lisetti and Fatma Nasoz. Using noninvasive wearable computers to recognize human emotions from physiological signals, 2004. doi:10.1155/S1110865704406192. 33

[193] Kun-Lin Liu, Wu-Jun Li, and Minyi Guo. Emoticon Smoothed Language Models for Twitter Sentiment Analysis. *Twenty-Sixth AAAI Conference on Artificial Intelligence*, jul 2012. URL: https://www.aaai.org/ocs/index.php/AAAI/AAAI12/paper/viewPaper/5083. 21

[194] Yong Jin Liu, Minjing Yu, Guozhen Zhao, Jinjing Song, Yan Ge, and Yuanchun Shi. Real-time movie-induced discrete emotion recognition from EEG signals. *IEEE Transactions on Affective Computing*, 2018. doi:10.1109/TAFFC.2017.2660485. 3, 4, 108

[195] Irene Lopatovska. Researching emotion: challenges and solutions. In *Proceedings of the 2011 iConference on - iConference '11*, New York, New York, USA, 2011. ACM Press. doi:10.1145/1940761. 141

[196] Ruth Luckasson, Sharon Borthwick-Duffy, Wil H. E. Buntinx, David L. Coulter, Ellis M. (Pat) Craig, Alya Reeve, Robert L. Schalock, Martha E. Snell, Deborah M. Spitalnik, Scott Spreat, and Marc J. Tassé. *Mental*

retardation: Definition, classification, and systems of supports. American Association on Mental Retardation, 10 edition, 2002. 52

[197] Diana MacLean, Asta Roseway, and Mary Czerwinski. MoodWings. In *Proceedings of the 6th International Conference on PErvasive Technologies Related to Assistive Environments - PETRA '13*, pages 1–8, New York, New York, USA, 2013. ACM Press. doi:10.1145/2504335.2504406. 43, 68

[198] Yuka Maeda, Masaki Sekine, and Toshiyo Tamura. The Advantages of Wearable Green Reflected Photoplethysmography. *Journal of Medical Systems*, 35(5):829–834, oct 2011. doi:10.1007/s10916-010-9506-z. 20

[199] Julia Makhaeva, Christopher Frauenberger, and Katharina Spiel. Creating creative spaces for co-designing with autistic children - The concept of a "Handlungsspielraum". In *ACM International Conference Proceeding Series*, 2016. doi:10.1145/2940299.2940306. 82

[200] Alexandru Ion Marinescu. Bach 2.0 - Generating classical music using recurrent neural networks. In *Procedia Computer Science*, 2019. doi:10.1016/j.procs.2019.09.166. xii, 25

[201] Justin Marley and Saeed Farooq. Mobile telephone apps in mental health practice: uses, opportunities and challenges. *BJPsych Bulletin*, 39(6):288–290, dec 2015. 16

[202] Hector P. Martinez, Yoshua Bengio, and Georgios Yannakakis. Learning deep physiological models of affect. *IEEE Computational Intelligence Magazine*, 8(2):20–33, 2013. 30, 32, 143

[203] R. Martinez, E. Irigoyen, A. Arruti, J. I. Martin, and J. Muguerza. A real-time stress classification system based on arousal analysis of the nervous system by an F-state machine. *Computer Methods and Programs in Biomedicine*, 2017. doi:10.1016/j.cmpb.2017.06.010. 33

[204] Mark Matthews and Gavin Doherty. In the Mood: Engaging Teenagers in Psychotherapy Using Mobile Phones. *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11*, 2011. 17

[205] Alban Maxhuni, Pablo Hernandez-Leal, L. Enrique Sucar, Venet Osmani, Eduardo F. Morales, and Oscar Mayora. Stress modelling and prediction in presence of scarce data. *Journal of Biomedical Informatics*, 63:344–356, oct 2016. 36

[206] Lisa J. Meltzer, Laura S. Hiruma, Kristin Avis, Hawley Montgomery-Downs, and Judith Valentin. Comparison of a commercial accelerometer with polysomnography and actigraphy in children and adolescents. *Sleep*, 2015. doi:10.5665/sleep.4918. 4

[207] Microsoft. Microsoft Band 2 features and functions, 2019. URL: support.microsoft.com/en-gb/help/4000313. 143

[208] Pardis Miri, Robert Flory, Andero Uusberg, Heather Culbertson, Richard H. Harvey, Agata Kelman, Davis Erik Peper, James J. Gross, Katherine Isbister, and Keith Marzullo. PIV: Placement, pattern, and personalization of an inconspicuous vibrotactile breathing pacer. *ACM Transactions on Computer-Human Interaction*, 2020. doi:10.1145/3365107. 41

[209] Varun Mishra. The Case for a Commodity Hardware Solution for Stress Detection. *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, pages 1717–1728, 2018. 48

[210] Alex J. Mitchell, Motahare Yadegarfar, John Gill, and Brendon Stubbs. Case finding and screening clinical utility of the Patient Health Questionnaire (PHQ-9 and PHQ-2) for depression in primary care: a diagnostic meta-analysis of 40 studies. *BJPsych Open*, 2(02):127–138, mar 2016. 12

[211] Changiz Mohiyeddini and Stuart Semple. Displacement behaviour regulates the experience of stress in men. *Stress*, 2013. 2, 80, 138, 161

[212] O. Hobart Mowrer. *Learning theory and behavior.* John Wiley Sons Inc, jan 1960. URL: /record/2005-06665-000, doi:10.1037/10802-000. 11

[213] Oscar Martinez Mozos, Virginia Sandulescu, Sally Andrews, David Ellis, Nicola Bellotto, Radu Dobrescu, and Jose Manuel Ferrandez. Stress de-

tection using wearable physiological and sociometric sensors. *International Journal of Neural Systems*, 2017. doi:10.1142/S0129065716500416. 12

[214] Amir Muaremi, Bert Arnrich, and Gerhard Tröster. Towards Measuring Stress with Smartphones and Wearable Devices During Workday and Sleep. *BioNanoScience*, 2013. doi:10.1007/s12668-013-0089-2. 37, 157

[215] Joe Mullenbach, Craig Shultz, J. Edward Colgate, and Anne Marie Piper. Exploring affective communication through variable-friction surface haptics. In *Conference on Human Factors in Computing Systems - Proceedings*, 2014. doi:10.1145/2556288.2557343. 41

[216] Inez Myin-Germeys, Zuzana Kasanova, Thomas Vaessen, Hugo Vachon, Olivia Kirtley, Wolfgang Viechtbauer, and Ulrich Reininghaus. Experience sampling methodology in mental health research: new insights and technical developments. *World Psychiatry*, 2018. 12

[217] Eitan Nahshoni, Dan Aravot, Dov Aizenberg, Mayanit Sigler, Gil Zalsman, Boris Strasberg, Shula Imbar, Edgar Adler, and Abraham Weizman. Heart Rate Variability in Patients with Major Depression. *Psychosomatics*, 2004. doi:10.1176/appi.psy.45.2.129. 20

[218] Benjamin W. Nelson and Nicholas B. Allen. Accuracy of consumer wearable heart rate measurement during an ecologically valid 24-hour period: Intraindividual validation study. *JMIR mHealth and uHealth*, 2019. doi:10.2196/10828. 4

[219] Hong Wei Ng, Viet Dung Nguyen, Vassilios Vonikakis, and Stefan Winkler. Deep learning for emotion recognition on small datasets using transfer learning. In *ICMI 2015 - Proceedings of the 2015 ACM International Conference on Multimodal Interaction*, pages 443–449, New York, New York, USA, nov 2015. Association for Computing Machinery, Inc. doi:10.1145/2818346.2830593. 36

[220] Karin Niemantsverdriet and Maarten Versteeg. Interactive Jewellery as Memory Cue. In *Proceedings of the TEI '16: Tenth International Conference on Tangible, Embedded, and Embodied Interaction - TEI '16*,

pages 532–538, New York, New York, USA, 2016. ACM Press. doi:10.1145/2839462.2856524. 17

[221] Robert P. Nolan, Markad V. Kamath, John S. Floras, Jill Stanley, Clement Pang, Peter Picton, and Quincy R. Young. Heart rate variability biofeedback as a behavioral neurocardiac intervention to enhance vagal heart rate control. *American Heart Journal*, 149(6):1137.e1–1137.e7, jun 2005. doi:10.1016/j.ahj.2005.03.015. 38

[222] Chris Noone and Michael J. Hogan. A randomised active-controlled trial to examine the effects of an online mindfulness intervention on executive control, critical thinking and key thinking dispositions in a university student sample. *BMC Psychology*, 6(1):13, dec 2018. 15

[223] Nvidia. NVIDIA Jetson Nano Developer Kit | NVIDIA Developer, 2019. URL: https://developer.nvidia.com/embedded/buy/jetson-nano-devkit. 88

[224] Cian O'Connor, Geraldine Fitzpatrick, Malcolm Buchannan-Dick, and James McKeown. Exploratory prototypes for video: Interpreting PD for a complexly disabled participant. In *ACM International Conference Proceeding Series*, 2006. doi:10.1145/1182475.1182500. 58

[225] Andreas Færøvig Olsen and Jim Torresen. Smartphone accelerometer data used for detecting human emotions. In *3rd International Conference on Systems and Informatics*, 2017. 21

[226] Jukka-Pekka Onnela and Scott L Rauch. Harnessing Smartphone-Based Digital Phenotyping to Enhance Behavioral and Mental Health. *Neuropsychopharmacology*, 41(7):1691–1696, jun 2016. URL: http://www.ncbi.nlm.nih.gov/pubmed/26818126http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4869063http://www.nature.com/articles/npp20167, doi:10.1038/npp.2016.7. 161

[227] Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. Learning and transferring mid-level image representations using convolutional neural net-

works. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2014. 35

[228] Sergio Oramas, Oriol Nieto, Francesco Barbieri, and Xavier Serra. Multi-label Music Genre Classification from Audio, Text, and Images Using Deep Features. jul 2017. 21

[229] Francisco Ordóñez and Daniel Roggen. Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition. *Sensors*, 16(1):115, jan 2016. doi:10.3390/s16010115. 30

[230] Xi Ouyang, Shigenori Kawaai, Ester Gue Hua Goh, Shengmei Shen, Wan Ding, Huaiping Ming, and Dong Yan Huang. Audio-visual emotion recognition using deep transfer learning and multiple temporal models. In *ICMI 2017 - Proceedings of the 19th ACM International Conference on Multimodal Interaction*, volume 2017-Janua, pages 577–582, New York, New York, USA, nov 2017. Association for Computing Machinery, Inc. doi:10.1145/3136755.3143012. 36

[231] Minna Pakanen, Ashley Colley, Jonna Häkkilä, Johan Kildal, and Vuokko Lantz. Squeezy bracelet. In *Proceedings of the 8th Nordic Conference on Human-Computer Interaction Fun, Fast, Foundational - NordiCHI '14*, pages 305–314, New York, New York, USA, 2014. ACM Press. doi:10.1145/2639189.2639238. 47

[232] N. Palmius, A. Tsanas, K. E. A. Saunders, A. C. Bilderbeck, J. R. Geddes, G. M. Goodwin, and M. De Vos. Detecting Bipolar Depression From Geographic Location Data. *IEEE Transactions on Biomedical Engineering*, 64(8):1761–1771, aug 2017. 31

[233] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning, 2010. doi:10.1109/TKDE.2009.191. 35, 118, 141

[234] Jonathan M. Peake, Graham Kerr, and John P. Sullivan. A critical review of consumer wearables, mobile applications, and equipment for providing biofeedback, monitoring stress, and sleep in physically active populations, 2018. doi:10.3389/fphys.2018.00743. 4

[235] Pearlmutter. Learning state space trajectories in recurrent neural networks. In *International Joint Conference on Neural Networks*, pages 365–372 vol.2. IEEE, 1989. doi:10.1109/IJCNN.1989.118724. 98

[236] Perkbox. THE 2O18 UK WORKPLACE STRESS SURVEY. Technical report, 2018. URL: https://www.perkbox.com/uk/resources/library/interactive-the-2018-uk-workplace-stress-survey. 1

[237] Amorette Perkins, Joseph Ridler, Daniel Browes, Guy Peryer, Caitlin Notley, and Corinna Hackmann. Experiencing mental health diagnosis: a systematic review of service user, clinician, and carer perspectives across clinical settings, 2018. doi:10.1016/S2215-0366(18)30095-6. 10

[238] R W Picard. *Affective Computing*. 1997. URL: http://www.media.mit.edu/~picard/. 2

[239] Rosalind W. Picard. Affective computing: challenges. *International Journal of Human-Computer Studies*, 59(1-2):55–64, jul 2003. doi:10.1016/S1071-5819(03)00052-1. 141

[240] Rosalind W. Picard, Elias Vyzas, and Jennifer Healey. Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(10):1175–1191, oct 2001. doi:10.1109/34.954607. 4, 118

[241] Kurt Plarre, Andrew Raij, Syed Monowar Hossain, Amin Ahsan Ali, Motohiro Nakajima, Mustafa Al'Absi, Emre Ertin, Thomas Kamarck, Santosh Kumar, Marcia Scott, Daniel Siewiorek, Asim Smailagic, and Lorentz E. Wittmers. Continuous inference of psychological stress from sensory measurements collected in the natural environment. In *Proceedings of the 10th ACM/IEEE International Conference on Information Processing in Sensor Networks, IPSN'11*, 2011. 12

[242] Thomas S Polzin and Alexander Waibel. EMOTION-SENSITIVE HUMAN-COMPUTER INTERFACES. 2000. URL: http://www.iscaaspeech.org/archive. 2

[243] Hugo F. Posada-Quintero, Jeffrey B. Bolkhovsky, Natasa Reljin, and Ki H. Chon. Sleep deprivation in young and healthy subjects is more sensitively identified by higher frequencies of electrodermal activity than by skin conductance level evaluated in the time domain. *Frontiers in Physiology*, 2017. doi:10.3389/fphys.2017.00409. 37

[244] Jacob Poushter. Smartphone Ownership and Internet Usage Continues to Climb in Emerging Economies. Technical report, 2016. 13

[245] Judith Proudfoot, Gordon Parker, Dusan Hadzi Pavlovic, Vijaya Manicavasagar, Einat Adler, and Alexis Whitton. Community attitudes to the appropriation of mobile phones for monitoring and managing depression, anxiety, and stress. *Journal of medical Internet research*, 12(5):e64, dec 2010. doi:10.2196/jmir.1475. 13

[246] Public Health England. Mental health of children in England, 2016. URL: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/575632/Mental_health_of_children_in_England.pdf. 166

[247] Rui Qiao, Chunmei Qing, Tong Zhang, Xiaofen Xing, and Xiangmin Xu. A novel deep-learning based framework for multi-subject emotion recognition. In *ICCSS 2017 - 2017 International Conference on Information, Cybernetics, and Computational Social Systems*, pages 181–185. IEEE, oct 2017. 33

[248] University of London Queen Mary. DEAP: A Dataset for Emotion Analysis using Physiological and Audiovisual Signals. *Ieee Trans. Affective Computing*, 2012. 3, 108

[249] Juan C. Quiroz, Min Hooi Yong, and Elena Geangu. Emotion-recognition using smart watch accelerometer data: Preliminary findings. In *2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers (UbiComp '17)*, 2017. 21

[250] U. Rajendra Acharya, K. Paul Joseph, N. Kannathal, Choo Min Lim, and Jasjit S. Suri. Heart rate variability: a review. *Medical Biological Engineering Computing*, 44(12):1031–1051, dec 2006. doi:10.1007/s11517-006-0119-0. 20

[251] Sreenivasan Ramasamy Ramamurthy and Nirmalya Roy. Recent trends in machine learning for human activity recognition—A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2018. doi:10.1002/widm.1254. 92

[252] Sarthak Singh Rawat, Abhishek Bisht, and Rahul Nijhawan. A Classifier Approach using Deep Learning for Human Activity Recognition. In *Proceedings of the IEEE International Conference Image Information Processing*, 2019. doi:10.1109/ICIIP47207.2019.8985926. 92

[253] Shireen L. Rizvi, Linda A. Dimeff, Julie Skutch, David Carroll, and Marsha M. Linehan. A Pilot Study of the DBT Coach: An Interactive Mobile Phone Application for Individuals With Borderline Personality Disorder and Substance Use Disorder. *Behavior Therapy*, 42(4):589–600, dec 2011. doi:10.1016/j.beth.2011.01.003. 16

[254] Roche. THE NHS AT 100, 2018. URL: http://roche-images.co.uk/design/corporate/rxukcomm01930b(1)a(1)_nhs100_roche_report.pdf. 14

[255] Paul A. Rodgers. Co-designing with people living with dementia. *CoDesign*, 2018. doi:10.1080/15710882.2017.1282527. 59

[256] Iyubanit Rodríguez, Gabriela Cajamarca, Valeria Herskovic, Carolina Fuentes, and Mauricio Campos. Helping Elderly Users Report Pain Levels: A Study of User Experience with Mobile and Wearable Interfaces. *Mobile Information Systems*, 2017:1–12, nov 2017. doi:10.1155/2017/9302328. 47

[257] A John Rush, Madhukar H Trivedi, Hicham M Ibrahim, Thomas J Carmody, Bruce Arnow, Daniel N Klein, John C Markowitz, Philip T Ninan, Susan Kornstein, Rachel Manber, Michael E Thase, James H Kocsis, and

Martin B Keller. The 16-Item Quick Inventory of Depressive Symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): a psychometric evaluation in patients with chronic major depression. *Biological psychiatry*, 54(5):573–83, sep 2003. 12

[258] Bryan C. Russell, Antonio Torralba, Kevin P. Murphy, and William T. Freeman. LabelMe: A database and web-based tool for image annotation. *International Journal of Computer Vision*, 2008. doi:10.1007/s11263-007-0090-8. 21, 88

[259] James A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178, dec 1980. doi:10.1037/h0077714. 11, 108

[260] Carolyn Saarni. Emotional development in childhood. *Encyclopedia on Early Childhood Develpment*, 2011. 77, 166

[261] Sara Sabour, Nicholas Frosst, and Geoffrey E. Hinton. Dynamic routing between capsules. In *Advances in Neural Information Processing Systems*, 2017. xii, 27

[262] Deepak Ranjan Sahoo, Timothy Neate, Yutaka Tokuda, Jennifer Pearson, Simon Robinson, Sriram Subramanian, Matt Jones, Deepak, Ranjan Sahoo, Jennifer Pear-Son, © Deepak, Ranjan Sahoo, and Jennifer Pear-Son. Tangible Drops: A Visio-Tactile Display Using Actuated Liquid-Metal Droplets. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, 2018. doi:10.1145/3173574.3173751. 48

[263] Michele R. Saloner and Maribeth Gettinger. Social inference skills in learning disabled and nondisabled children. *Psychology in the Schools*, 1985. doi:10.1002/1520-6807(198504)22:2<201::AID-PITS2310220214>3.0.CO;2-V. 53

[264] Pedro Sanches, Kristina Höök, Elsa Vaara, Claus Weymann, Markus Bylund, Pedro Ferreira, Nathalie Peira, and Marie Sjölinder. Mind the body! Designing a mobile stress management application encouraging personal reflection. In *DIS 2010 - Proceedings of the 8th ACM Conference on Designing Interactive Systems*, 2010. doi:10.1145/1858171.1858182. 42

[265] Pedro Sanches, Axel Janson, Pavel Karpashevich, Camille Nadal, Chengcheng Qu, Claudia Daudén Roquet, Muhammad Umair, Charles Windlin, Gavin Doherty, Kristina Höök, and Corina Sas. HCI and Affective Health Taking stock of a decade of studies and charting future research directions. In *Conference on Human Factors in Computing Systems - Proceedings*, page 17. Association for Computing Machinery, may 2019. doi:10.1145/3290605.3300475. 57

[266] E Sanders. Design Research in 2006. *Design Research Society*, 2006. 54

[267] E. B.-N. Sanders. Generative Tools for Co-designing. In *Collaborative Design*, pages 3–12. Springer London, 2000. doi:10.1007/978-1-4471-0779-8_1. 59

[268] Elizabeth B. N. Sanders and Uday Dandavate. Design for Experiencing: New tools. In *First International Conference on Design and Emotion*, 1999. 59, 60

[269] Elizabeth B.-N. Sanders and Pieter Jan Stappers. Co-creation and the new landscapes of design. *CoDesign*, 2008. doi:10.1080/15710880701875068. 54

[270] Akane Sano and Rosalind W Picard. Stress Recognition using Wearable Sensors and Mobile Phones. *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, 2013. 31

[271] Allah Bux Sargano, Xiaofeng Wang, Plamen Angelov, and Zulfiqar Habib. Human action recognition using transfer learning with deep representations. In *Proceedings of the International Joint Conference on Neural Networks*, volume 2017, pages 463–469. IEEE, jun 2017. doi:10.1109/IJCNN.2017.7965890. 35

[272] Martha Patricia Sarmiento-Pelayo. Co-design: A central approach to the inclusion of people with disabilities. *Revista Facultad de Medicina*, 2015. doi:10.15446/revfacmed.v63n3sup.49345. 54

[273] Federico Sarzotti. Self-Monitoring of Emotions and Mood Using a Tangible Approach. *Computers*, 7(1):7, jan 2018. 17, 47, 57, 88

[274] Robert L. Schalock, Sharon A. Borthwick-Duffy, Valerie J. Bradley, Wil H. E. Buntinx, David L. Coulter, Ellis M. Craig, Sharon C. Gomez, Yves Lachapelle, Ruth Luckasson, Alya Reeve, Karrie A. Shogren, Martha E. Snell, Scott Spreat, Marc J. Tasse, James R. Thompson, Miguel A. Verdugo-Alonso, Michael L. Wehmeyer, and Mark H. Yeager. *Intellectual Disability: Definition, Classification, and Systems of Supports. Eleventh Edition.* 2010. 52

[275] Björn Schuller, R A Calvo, S Dmello, J Gratch, and A Kappas. Multimodal affect databases: Collection, challenges, and chances. *Handbook of Affective Computing*, pages 323–333, 2015. 141

[276] Johannes Schumm, Eth Zurich, Ulrike Ehlert, Cornelia Setz, Bert Arnrich, Roberto La Marca, and Gerhard Tröster. Discriminating stress from cognitive load using a wearable EDA device. *IEEE Transactions on Information Technology in Biomedicine*, 14(2), 2010. 2, 20, 32

[277] Hans Selye. Stress without Distress. In George Serban, editor, *Serban G. (eds). Psychopathology of Human Adaptation*, pages 137–146. Springer US, 1976. doi:10.1007/978-1-4684-2238-2_9. 12

[278] Ssang-Hee Seo and Jung-Tae Lee. Stress and EEG. In *Convergence and Hybrid Information Technologies*. 2012. arXiv:9809069v1, doi:10.5772/9651. 41

[279] Joan Serrà, Santiago Pascual, and Alexandros Karatzoglou. Towards a Universal Neural Network Encoder for Time Series. In *Frontiers in Artificial Intelligence and Applications*, 2018. doi:10.3233/978-1-61499-918-8-120. 28

[280] Monit Shah Singh, Vinaychandran Pondenkandath, Bo Zhou, Paul Lukowicz, Marcus Liwicki, and Tu Kaiserslautern. Transforming Sensor Data to the Image Domain for Deep Learning-an Application to Footstep Detection. arXiv:1701.01077v3, doi:10.1109/IJCNN.2017.7966182. 35

[281] Nandita Sharma and Tom Gedeon. Objective measures, sensors and computational techniques for stress recognition and classification: A survey.

Computer Methods and Programs in Biomedicine, 108(3):1287–1301, dec 2012. 2, 32

[282] Takanori Shibata and Kazuyoshi Wada. Robot Therapy: A New Approach for Mental Healthcare of the Elderly – A Mini-Review. *Gerontology*, 57(4):378–386, 2011. doi:10.1159/000319015. 44

[283] Benjamin Shickel, Martin Heesacker, Sherry Benton, and Parisa Rashidi. Hashtag Healthcare: From Tweets to Mental Health Journals Using Deep Transfer Learning. aug 2017. 37

[284] Saul Shiffman, Arthur A Stone, and Michael R Hufford. Ecological momentary assessment. *Annual review of clinical psychology*, 4:1–32, 2008. 45, 88

[285] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015. arXiv:1409.1556. 35

[286] Laurianne Sitbon and Shanjana Farhin. Co-designing interactive applications with adults with intellectual disability: A case study. In *ACM International Conference Proceeding Series*, 2017. doi:10.1145/3152771.3156163. 55

[287] Iouliia Skliarova and Valery Sklyarov. Hardware/software co-design. In *Lecture Notes in Electrical Engineering*. 2019. doi:10.1007/978-3-030-20721-2_6. 54

[288] Wina Smeenk, Janienke Sturm, and Berry Eggen. Empathic handover: how would you feel? Handing over dementia experiences and feelings in empathic co-design. *CoDesign*, 2018. doi:10.1080/15710882.2017.1301960. 59

[289] Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic. A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing*, 2012. doi:10.1109/T-AFFC.2011.25. 3, 108

[290] Tengfei Song, Wenming Zheng, Peng Song, and Zhen Cui. EEG Emotion Recognition Using Dynamical Graph Convolutional Neural Networks. *IEEE Transactions on Affective Computing*, 2018. doi:10.1109/TAFFC.2018.2817622. 3, 33, 108

[291] Herbert Spencer González, Vanessa Vega Córdova, Katherine Exss Cid, Marcela Jarpa Azagra, and Izaskun Álvarez-Aguado. Including intellectual disability in participatory design processes: Methodological adaptations and supports. In *ACM International Conference Proceeding Series*, 2020. doi:10.1145/3385010.3385023. 57, 82, 86

[292] Katta Spiel. SIGCHI outstanding dissertation award: Evaluating experiences of autistic children with technologies in co-design. In *Conference on Human Factors in Computing Systems - Proceedings*, pages 1–4, New York, NY, USA, apr 2020. Association for Computing Machinery. URL: https://dl.acm.org/doi/10.1145/3334480.3386151, doi:10.1145/3334480.3386151. 55

[293] Katta Spiel, Emeline Brulé, Christopher Frauenberger, Gilles Bailly, and Geraldine Fitzpatrick. Micro-ethics for participatory design with marginalised children. 2018. doi:10.1145/3210586.3210603. 55

[294] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, and Ruslan Salakhutdinov. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Technical report, 2014. 126

[295] Marc Steen. Co-design as a process of joint inquiry and imagination. *Design Issues*, 2013. doi:10.1162/DESI_a_00207. 54

[296] Marc Steen, Menno Manschot, and Nicole de Koning. Benefits of co-design in service design projects. *International Journal of Design*, 2011. 54

[297] Arthur A Stone, Saul Shiffman, Joseph E Schwartz, Joan E Broderick, and Michael R Hufford. Patient non-compliance with paper diaries. *BMJ (Clinical research ed.)*, 324(7347):1193–1194, may 2002. 12, 13

[298] Arthur A Stone, Jaylan S Turkkan, Christine A Bachrach, Jared B Jobe, Howard S Kurtzman, and Virginia S Cain, editors. *The science of self-*

report: Implications for research and practice. Lawrence Erlbaum Associates Publishers, Mahwah, NJ, US, 2000. 12

[299] Feng-Tso Sun, Cynthia Kuo, Heng-Tze Cheng, Senaka Buthpitiya, Patricia Collins, and Martin Griss. Activity-Aware Mental Stress Detection Using Physiological Sensors. pages 211–230. Springer, Berlin, Heidelberg, oct 2012. doi:10.1007/978-3-642-29336-8_12. 48

[300] Si Sun, Marion J Ball, and Ching-hua Chen. Investigating the Role of Context in Perceived Stress Detection in the Wild. *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing*, (July 2017):1708–1716, 2018. 32

[301] Karush Suri and R. Gupta. Continuous sign language recognition from wearable IMUs using deep capsule networks and game theory. *Computers and Electrical Engineering*, 2019. 27

[302] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A. Alemi. Inception-v4, inception-ResNet and the impact of residual connections on learning. In *31st AAAI Conference on Artificial Intelligence, AAAI 2017*, 2017. arXiv:1602.07261. 29

[303] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the Inception Architecture for Computer Vision. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016. arXiv:1512.00567, doi:10.1109/CVPR.2016.308. 35

[304] Joachim Taelman, S. Vandeput, A. Spaepen, and S. Van Huffel. Influence of mental stress on heart rate and heart rate variability. In *IFMBE Proceedings*, volume 22, pages 1366–1369. Springer, Berlin, Heidelberg, 2008. doi:10.1007/978-3-540-89208-3_324. 145

[305] Mohammed Taj-Eldin, Christian Ryan, Brendan O'flynn, and Paul Galvin. A review of wearable solutions for physiological and emotional monitoring for use by people with autism spectrum disorder and their caregivers, 2018. doi:10.3390/s18124271. 4

[306] Masahiro Tanida, Masako Katsuyama, and Kaoru Sakatani. Relation between mental stress-induced prefrontal cortex activity and skin conditions: A near-infrared spectroscopy study. *Brain Research*, 1184:210–216, dec 2007. 20

[307] Pattreeya Tanisaro and Gunther Heidemann. Time series classification using time warping invariant Echo State Networks. In *Proceedings - 2016 15th IEEE International Conference on Machine Learning and Applications, ICMLA 2016*, 2017. doi:10.1109/ICMLA.2016.166. 28

[308] Emmanuel Munguia Tapia, Stephen S. Intille, William Haskell, Kent Larson, Julie Wright, Abby King, and Robert Friedman. Real-time recognition of physical activities and their intensities using wireless accelerometers and a heart rate monitor. In *Proceedings - International Symposium on Wearable Computers, ISWC*, 2007. doi:10.1109/ISWC.2007.4373774. 21

[309] The International Well Being Group. *Personal Wellbeing Index – Adult (English)*. 2006. 108

[310] The Mental Health Foundation. Fundamental facts about mental health. Technical report, 2015. 1

[311] Tatiana Tommasi, Francesco Orabona, and Barbara Caputo. Safety in numbers: Learning categories from few examples with multi model knowledge transfer. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010. doi:10.1109/CVPR.2010.5540064. 35

[312] John Torous, Rohn Friedman, and Matcheri Keshavan. Smartphone ownership and interest in mobile applications to monitor symptoms of mental health conditions. *JMIR mHealth and uHealth*, 2(1):e2, jan 2014. doi:10.2196/mhealth.2994. 13

[313] Sarah Geiss Trusz, Amy W. Wagner, Joan Russo, Jeff Love, and Douglas F. Zatzick. Assessing Barriers to Care and Readiness for Cognitive Behavioral Therapy in Early Acute Care PTSD Interventions. *Psychiatry: Interpersonal and Biological Processes*, 74(3):207–223, sep 2011. doi:10.1521/psyc.2011.74.3.207. 13

[314] A. Tsanas, K.E.A. Saunders, A.C. Bilderbeck, N. Palmius, M. Osipov, G.D. Clifford, G.M. Goodwin, and M. De Vos. Daily longitudinal self-monitoring of mood variability in bipolar disorder and borderline personality disorder. *Journal of Affective Disorders*, 205:225–233, nov 2016. 14, 158

[315] Theodora Tsikrika, Christos Diou, Arjen P. de Vries, and Anastasios Delopoulos. Image annotation using clickthrough data. 2009. doi:10.1145/1646396.1646415. 21, 88

[316] G. Tur, R.E. Schapire, and D. Hakkani-Tur. Active learning for spoken language understanding. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, volume 1, pages I–276–I–279. IEEE, 2003. 21

[317] B. Ullmer and H. Ishii. Emerging frameworks for tangible user interfaces. *IBM Systems Journal*, 2000. doi:10.1147/sj.393.0915. 2, 52

[318] Muhammad Umair, Muhammad Hamza Latif, and Corina Sas. Dynamic displays at wrist for real time visualization of affective data. In *DIS 2018 - Companion Publication of the 2018 Designing Interactive Systems Conference*, 2018. doi:10.1145/3197391.3205436. 42

[319] Terumi Umematsu, Akane Sano, Sara Taylor, and Rosalind W. Picard. Improving Students' Daily Life Stress Forecasting using LSTM Neural Networks. pages 1–4. Institute of Electrical and Electronics Engineers (IEEE), sep 2019. 34

[320] Evgeny G. Vaschillo, Bronya Vaschillo, and Paul M. Lehrer. Characteristics of resonance in heart rate variability stimulated by biofeedback. *Applied Psychophysiology Biofeedback*, 2006. doi:10.1007/s10484-006-9009-3. 20

[321] Cati Vaucelle, Leonardo Bonanni, and Hiroshi Ishii. Design of haptic interfaces for therapy. In *Proceedings of the 27th international conference on Human factors in computing systems - CHI 09*, page 467, New York, New York, USA, 2009. ACM Press. doi:10.1145/1518701.1518776. 40

[322] Jenn Wortman Vaughan. Making Better Use of the Crowd: How Crowdsourcing Can Advance Machine Learning Research. *JMLR*, 18, jan 2019. URL: https://www.microsoft.com/en-us/research/publication/making-better-use-of-the-crowd-how-crowdsourcing-can-advance-machine-learn 21, 88

[323] Daniela Villani, Alessandra Grassi, Chiara Cognetta, Pietro Cipresso, Davide Toniolo, and Giuseppe Riva. The effects of a mobile stress management protocol on nurses working with cancer patients: a preliminary controlled study. In *MMVR*, pages 524–528, 2012. 15

[324] Daniela Villani, Alessandra Grassi, Chiara Cognetta, Davide Toniolo, Pietro Cipresso, and Giuseppe Riva. Self-help stress management training through mobile phones: An experience with oncology nurses. *Psychological Services*, 10(3):315–322, 2013. 15

[325] John Vines, Rachel Clarke, Peter Wright, John McCarthy, and Patrick Olivier. Configuring participation: On how we involve people in design. In *Conference on Human Factors in Computing Systems - Proceedings*, 2013. doi:10.1145/2470654.2470716. 54

[326] Froukje Sleeswijk Visser. *Bringing the everyday life of people into design*. 2009. doi:Isbn978-90-9024244-6. 54

[327] Eric von Hippel. “Sticky Information” and the Locus of Problem Solving: Implications for Innovation. *Management Science*, 1994. doi:10.1287/mnsc.40.4.429. 54

[328] O. F. Wahl. Mental Health Consumers’ Experience of Stigma. *Schizophrenia Bulletin*, 25(3):467–478, jan 1999. doi:10.1093/oxfordjournals.schbul.a033394. 13

[329] Tracey Wallace, John T Morris, Scott Bradshaw, and Corissa Bayer. BreatheWell: Developing a Stress Management App on Wearables for TBI PTSD. *Journal on Technology and Persons with Disabilities Santiago, J*, 2017. 31

[330] Jan Walmsley. Inclusive learning disability research: The (nondisabled) researcher's role. *British Journal of Learning Disabilities*, 2004. doi:10.1111/j.1468-3156.2004.00281.x. 56

[331] Jan Walmsley, Iva Strnadová, and Kelley Johnson. The added value of inclusive research. *Journal of Applied Research in Intellectual Disabilities*, 2018. doi:10.1111/jar.12431. 56

[332] C. L. Walters. The Psychological and Physiological Effects of Vibrotactile Stimulation, Via a Somatron, on Patients Awaiting Scheduled Gynecological Surgery. *Journal of Music Therapy*, 33(4):261–287, dec 1996. 40

[333] Dong Wang and Thomas Fang Zheng. Transfer learning for speech and language processing. In *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC 2015*, pages 1225–1237. Institute of Electrical and Electronics Engineers Inc., feb 2016. 35

[334] Ri Wang, Maysum Panju, and Mahmood Gohari. Classification-based RNN machine translation using GRUs, 2017. arXiv:1703.07841. xii, 26

[335] 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. arXiv:1803.08955, doi:10.1145/3191775. 31

[336] Zhiguang Wang and Tim Oates. Imaging time-series to improve classification and imputation. In *IJCAI International Joint Conference on Artificial Intelligence*, 2015. arXiv:1506.00327. 19, 141, 147, 148

[337] Zhiguang Wang, Weizhong Yan, and Tim Oates. Time series classification from scratch with deep neural networks: A strong baseline. In *Proceedings of the International Joint Conference on Neural Networks*, 2017. arXiv:1611.06455, doi:10.1109/IJCNN.2017.7966039. 28

[338] B Weiner and S Graham. An attributional approach to emotional development. In *Emotions, cognition, and behavior*. 1985. 11

[339] John R. Weisz, Irwin N. Sandler, Joseph A. Durlak, and Barry S. Anton. Promoting and Protecting Youth Mental Health Through Evidence-Based Prevention and Treatment. *American Psychologist*, 60(6):628–648, 2005. doi:10.1037/0003-066X.60.6.628. 12, 13

[340] Wanhui Wen, Guangyuan Liu, Nanpu Cheng, Jie Wei, Pengchao Shangguan, and Wenjin Huang. Emotion recognition based on multi-variant correlation of physiological signals. *IEEE Transactions on Affective Computing*, 2014. doi:10.1109/TAFFC.2014.2327617. 4, 32, 118

[341] Amanda L. Wheat and Kevin T. Larkin. Biofeedback of Heart Rate Variability and Related Physiology: A Critical Review. *Applied Psychophysiology and Biofeedback*, 35(3):229–242, sep 2010. doi:10.1007/s10484-010-9133-y. 39

[342] Joseph Wherton, Paul Sugarhood, Rob Procter, Sue Hinder, and Trisha Greenhalgh. Co-production in practice: How people with assisted living needs can help design and evolve technologies and services. *Implementation Science*, 2015. doi:10.1186/s13012-015-0271-8. 55

[343] WHO and World Health Organisation. WHO | Mental health: a state of well-being, 2014. URL: https://www.who.int/features/factfiles/mental_health/en/. 1, 11

[344] J. Wijsman, B. Grundlehner, Hao Liu, H. Hermens, and J. Penders. Towards mental stress detection using wearable physiological sensors. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 1798–1801. IEEE, aug 2011. 2, 34

[345] Robin Williams. Why is it difficult to achieve e-health systems at scale? *Information, Communication Society*, 19(4):540–550, apr 2016. 46

[346] Charles Windlin. Designing with the body: Addressing emotion regulation and expression. In *DIS 2020 Companion - Companion Publication*

of the 2020 ACM Designing Interactive Systems Conference, pages 557–562, New York, NY, USA, jul 2020. Association for Computing Machinery, Inc. URL: https://dl.acm.org/doi/10.1145/3393914.3395835, doi:10.1145/3393914.3395835. 42

[347] Kieran Woodward and Eiman Kanjo. iFidgetCube: Tangible Fidgeting Interfaces (TFIs) to Monitor and Improve Mental Wellbeing. *IEEE Sensors Journal*, pages 1–1, 2020. URL: https://ieeexplore.ieee.org/document/9223707/, doi:10.1109/JSEN.2020.3031163. 10, 117, 160

[348] Kieran Woodward, Eiman Kanjo, David Brown, T. M. McGinnity, Becky Inkster, Donald J Macintyre, and Athanasios Tsanas. Beyond Mobile Apps: A Survey of Technologies for Mental Well-being. *IEEE Transactions on Affective Computing*, 2020. 10

[349] Kieran Woodward, Eiman Kanjo, David J Brown, and Becky Inkster. TangToys: Smart Toys to Communicate and Improve Children's Wellbeing. In *Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers*, New York, NY, USA, 2020. ACM. URL: https://doi.org/10.1145/3410530.3414375. 160

[350] Kieran Woodward, Eiman Kanjo, David J. Brown, and T.M. McGinnity. Towards Personalised Mental Wellbeing Recognition On-Device using Transfer Learning "in the Wild". In *IEEE International Smart Cities Conference 2021*, 2021. 117

[351] Kieran Woodward, Eiman Kanjo, Andreas Oikonomou, and Samuel Burton. Emoecho: A tangible interface to convey and communicate emotions. In *UbiComp/ISWC 2018 - 2018 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2018 ACM International Symposium on Wearable Computers*, 2018. 160

[352] Kieran Woodward, Eiman Kanjo, Andreas Oikonomou, and Alan Chamberlain. LabelSens: enabling real-time sensor data labelling at the point of collection using an artificial intelligence-based approach. *Personal and Ubiquitous Computing*, 24(5):709–722, jun 2020. 87

[353] World Health Organization. *The ICD-10 classification of mental and behavioural disorders: Diagnostic criteria for research*. 1993. 13

[354] Guanghua Wu, Guangyuan Liu, and Min Hao. The analysis of emotion recognition from GSR based on PSO. In *Proceedings - 2010 International Symposium on Intelligence Information Processing and Trusted Computing, IPTC 2010*, 2010. doi:10.1109/IPTC.2010.60. 32

[355] Mike Wu, Brian Richards, and Ron Baecker. Participatory design with individuals who have amnesia. In *PDC 2004 - Proceedings of the Eight Participatory Design Conference 2004 - Artful Integration: Interweaving Media, Materials and Practices*, 2004. doi:10.1145/1011870.1011895. 59

[356] Xiaofen Xing, Zhenqi Li, Tianyuan Xu, Lin Shu, Bin Hu, and Xiangmin Xu. SAE+LSTM: A new framework for emotion recognition from multi-channel EEG. *Frontiers in Neurorobotics*, 13, 2019. 34

[357] Chen Yi Chao Lung Yang, Chen Yi Chao Lung Yang, Zhi Xuan Chen, and Nai Wei Lo. Multivariate Time Series Data Transformation for Convolutional Neural Network. In *Proceedings of the 2019 IEEE/SICE International Symposium on System Integration, SII 2019*, pages 188–192. Institute of Electrical and Electronics Engineers Inc., apr 2019. doi:10.1109/SII.2019.8700425. 19, 141, 147

[358] Wenpeng Yin, Katharina Kann, Mo Yu, and Hinrich Schütze. Comparative Study of CNN and RNN for Natural Language Processing. feb 2017. URL: http://arxiv.org/abs/1702.01923, arXiv:1702.01923. 26

[359] Hayley A. Young and David Benton. Heart-rate variability: A biomarker to study the influence of nutrition on physiological and psychological health?, 2018. doi:10.1097/FBP.0000000000000383. 37

[360] Eman M. G. Younis, Eiman Kanjo, and Alan Chamberlain. Designing and evaluating mobile self-reporting techniques: crowdsourcing for citizen science. *Personal and Ubiquitous Computing*, pages 1–10, mar 2019. 13, 21, 22, 88

[361] Xin Yu and Panos D. Prevedouros. Performance and Challenges in Utilizing Non-Intrusive Sensors for Traffic Data Collection. *Advances in Remote Sensing*, 2013. doi:10.4236/ars.2013.22006. 141

[362] Roberto Zangróniz, Arturo Martínez-Rodrigo, José Manuel Pastor, María T López, and Antonio Fernández-Caballero. Electrodermal Activity Sensor for Classification of Calm/Distress Condition. *Sensors (Basel, Switzerland)*, 17(10), oct 2017. 2

[363] Alexandros Zenonos, Aftab Khan, Georgios Kalogridis, Stefanos Vatsikas, Tim Lewis, and Mahesh Sooriyabandara. HealthyOffice: Mood recognition at work using smartphones and wearable sensors. In *2016 IEEE International Conference on Pervasive Computing and Communication Workshops, PerCom Workshops 2016*, 2016. doi:10.1109/PERCOMW.2016.7457166. 38, 157

[364] Ran Zhang, Hongyang Tao, Lifeng Wu, and Yong Guan. Transfer Learning with Neural Networks for Bearing Fault Diagnosis in Changing Working Conditions. *IEEE Access*, 5:14347–14357, 2017. 23

[365] Xiao Zhang, Wenzhong Li, Xu Chen, and Sanglu Lu. MoodExplorer: Towards Compound Emotion Detection via Smartphone Sensing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. Article*, 1(176), 2017. doi:10.1145/3161414. 30

[366] Xiao Zhang, Yongqiang Lyu, Xiaomin Luo, Jingyu Zhang, Chun Yu, Hao Yin, and Yuanchun Shi. Touch Sense. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(2):1–18, 2018. doi:10.1145/3214290. 30

[367] Zhan Zhang, Yufei Song, Liqing Cui, Xiaoqian Liu, and Tingshao Zhu. Emotion recognition based on customized smart bracelet with built-in accelerometer. *PeerJ*, 4:e2258, 2016. 21

[368] Wei-Long Zheng and Bao-Liang Lu. Personalizing EEG-Based Affective Models with Transfer Learning. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16)*, 2016. 3, 36, 108

[369] Yi Zheng, Qi Liu, Enhong Chen, Yong Ge, and J. Leon Zhao. Time series classification using multi-channels deep convolutional neural networks. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2014. doi:10.1007/978-3-319-08010-9_33. 29

[370] Jan Henrik Ziegeldorf, Oscar Garcia Morchon, and Klaus Wehrle. Privacy in the internet of things: Threats and challenges. *Security and Communication Networks*, 2014. doi:10.1002/sec.795. 141

[371] F R H Zijlstra and L Doorn. The construction of a scale to measure subjective effort. *Delft, The Netherlands: Delft University of Technology, Department of Philosophy and Social Sciences*, 1985. 120

[372] Muhammad Zubair, Changwoo Yoon, Hyunyoung Kim, Jisu Kim, and Jinsul Kim. Smart wearable band for stress detection. In *2015 5th International Conference on IT Convergence and Security, ICITCS 2015 - Proceedings*, 2015. doi:10.1109/ICITCS.2015.7293017. 32