Univerza v Ljubljani
Fakulteta za računalništvo in informatiko

Tilen Matkovič

Ugotavljanje kognitivne obremenjenosti človeka z uporabo brezžičnih signalov

magistrsko delo

magistrski program druge stopnje
računalništvo in informatika

Mentor: doc. dr. Veljko Pejović

Ljubljana, 2018

University of Ljubljana
Faculty of Computer and Information Science

Tilen Matkovič

Inferring cognitive load using wireless signals

master's thesis

the 2nd cycle master's study programme
computer and information science

Supervisor: doc. dr. Veljko Pejović

Ljubljana, 2018

Contents

List of used acronmys

acronym	meaning
HCI	human-computer interaction
TLX	task load index
GSR	galvanic skin response
ECG	electrocardiogram
EEG	electroencephalography
HRV	heart rate variability
BCG	ballistocardiography
CSI	channel state information
UWB	ultra-wideband
IR	impulse-radio
RF	radio frequency
CPU	central processing unit
SDR	software defined radio
FMCW	frequency modulated carrier wave
USRP	universal software radio peripheral
RRV	respiratory rate variability
FFT	fast Fourier transform
LS	Lomb-Scargle

Povzetek

Naslov: Ugotavljanje kognitivne obremenjenosti človeka z uporabo brezžičnih signalov

Z vsestranskim računalništvom in njegovo interakcijo z ljudmi bi lahko izboljšali uporabniško izkušnjo, če bi naprave lahko razbrale kognitivno obremenjenost uporabnikov. Trenutni načini ocenjevanja kognitivne obremenitve človeka so, z nekaj izjemami, zasnovani na metodah, ki zahtevajo fizični stik merilnih oprem in uporabnika. V tem delu predstavimo sistem Wi-Mind za ocenjevanje kognitivne obremenjenosti človeka z uporabo brezžičnih signalov. Wi-Mind temelji na programsko definiranem radijskem radarju, ki meri zelo majhne gibe človeka, ki so rezultat dihanja in srčnega utripa. Le-to nam omogoča ocenjevanje kognitivne obremenjenosti osebe. Sistem smo preizkusili in testirali na triindvajsetih prostovoljcih, ki so reševali naloge različnih težavnosti. Rezultati kažejo, da Wi-Mind do neke mere lahko ugotovi ali se oseba ukvarja z reševanjem naloge. Ocenjevanje direktne kognitivne obremenjenosti, s katero bi lahko ugotovili uporabnikovo zagnanost v problem, ostaja izziv.

Ključne besede

življenjski znaki, procesiranje signalov, podatkovno rudarjenje, kognitivna obremenjenost, brezžično zaznavanje, programsko definirani radio

Abstract

Title: Inferring cognitive load using wireless signals

From not disturbing a focused programmer, to entertaining a restless commuter waiting for a train, ubiquitous computing devices could greatly enhance their interaction with humans, should these devices only be aware of the user's cognitive load. However, current means of assessing cognitive load are, with a few exceptions, based on intrusive methods requiring physical contact of the measurement equipment and the user. In this thesis we propose Wi-Mind, a system for remote cognitive load assessment through wireless sensing. Wi-Mind is based on a software-defined radio-based radar that measures sub-millimeter movements related to a person's breathing and heartbeats, which, in turn allow us to infer the person's cognitive load. We built the system and tested it with 23 volunteers being engaged in different tasks. Results show that while Wi-Mind manges to detect whether one is engaged in a cognitively demanding task, the inference of the exact cognitive load level remains challenging.

Keywords

vital signs, signal processing, data mining, cognitive load, wireless sensing, software-defined radio

Razširjeni povzetek

I Uvod in sorodna dela

Z napravami, s katerimi opravljamo dnevna opravila, delamo tako koristne opravke, kot tudi tiste, ki nas zmotijo in odvrnejo od pozornosti pri trenutni aktivnosti. Razna opozorila na telefonu in podobne motnje, ki pogostokrat prekinejo naše trenutno delo, lahko privedejo do nezbranosti, slabšega učinka pri delu [1] in k stresu [2]. Prekinljivost pri delu bi lahko zmanjšali s tem, da bi tovrstne naprave avtomatično prepoznale trenutno uporabniško aktivnost oziroma kognitivno obremenjenost človeka. Kognitivna obremenjenost se lahko ugotovi s subjektivno samooceno, ki se reši po opravljeni nalogi (na primer NASA-TLX testi [3]). Drug način za oceno kognitivne obremenjenosti, ki pa je bolj primeren za uporabo v realnem času, je merjenje fizioloških signalov, ki korelirajo z mentalnim naporom in so rezultat avtonomnega živčevja ter kardiovaskularnega sistema. Signale teh sistemov se lahko zaznava z meritvijo srčnega utripa [4], dihanja [5], aktivnosti v možganih [6], velikostjo zenic in s podobnimi signali, ki jih zavestno težko kontroliramo. Večino teh signalov lahko merimo z opremo, ki zahteva fizični stik naprave z osebo.

Razvoj tehnologije je privedel do zaznavanja življenjskih znakov, ki korelirajo s kognitivno obremenjenostjo, brez kakršnegakoli stika med človekom in napravo. Bodisi so to pristopi, ki izkoriščajo kamero [7, 8], bodisi pristopi s katerimi s pošiljanjem brezžičnih signalov lahko ugotovimo dihanje in srčni utrip človeka [9, 10, 11, 12, 13, 14, 15]. V našem delu se osredotočimo

na Vital-Radio [9], ki omogoča merjenje minimalnih gibov človeka, ki so rezultat dihanja in srčenga utripa. Vital-Radio deluje z radarsko FMCW (angl. frequency modulated carrier wave) metodo. Z omenjeno tehnologijo, ki brezžično zajema življenjske znake uporabnika, bomo raziskali možnost ugotavljanja trenutne kognitivne obremenjenosti osebe.

II Wi-Mind sistem

Wi-Mind sistem je zgrajen na predpostavki, da življenjski znaki, kot so dihanje in srčni utrip, korelirajo s človekovo kognitivno obremenjenostjo. Naša implementacija vsebuje dva glavna modula, in sicer *modul za brezžično zajemanje signala* in *modul strojnega učenja*. Ideja pri *modulu za brezžično zajemanje signala* je osnovana na že omenjenem Vital-Radiu [9], ki uporablja radarsko FMCW tehnologijo. Omenjeni radar se v našem primeru poganja na programsko definiranem radiu (angl. *software defined radio*) – konceptu, ki omogoča visoko fleksibilnost pošiljanja/zajemanja brezžičnih signalov z ustrezno strojno in programsko opremo na navadnem osebnem računalniku. Radar zajema signal, ki potuje do človeka (primarno do prs), se odbije in potuje nazaj do sprejemne antene – na ta način vidimo razdaljo, ki jo prepotuje signal. S periodičnimi minimalnimi gibi človeka (dihanjem in srčnim utripom – pojav, ki nam omogoča brezžično zajemanje srčnega utripa, se imenuje balistokardiografija) lahko nato izločimo koristne informacije in jih uporabimo v algoritmih strojnega učenja.

Surov signal razbijemo na časovna okna ter na vsakem od teh oken izvedemo filtriranje oziroma izločevanje značilk. Ker je radar zelo občutljiv, lahko vsebuje veliko šuma in posledično negativno vpliva na rezultat, je dobro tovrstne šume izločiti. Šum je lahko posledica hitrih gibov (na primer premik roke) ali zelo počasnega gibanja, ki se jih da izločiti s pasovnim filtrom (za dihanje uporabimo pasovni filter med 0.083 Hz in 1 Hz). Nato iz signala dobimo frekvenco dihanja, spremembo frekvence dihanja, energijo v posameznih frekvenčnih pasovih in variabilnost dihanja. Na podoben prin-

cip lahko dobimo informacije o bitju srca – tokrat uporabimo pasovni filter med 0.83 Hz in 2.5 Hz. S pretvorbo časovnega okna v frekvenčni prostor nato izločimo frekvenco srčnega utripa, z merjenjem posameznih utripov v časovnem prostoru pa lahko dobimo variabilnost srčnega utripa (HRV, angl. *heart rate variability*). Le-ta je posledica avtonomnega živčevja, ki regulira podzavestne telesne funkcije.

Z vsemi potencialno koristnimi informacijami nato zaženemo algoritme strojnega učenja, ki ocenijo trenutno kognitivno stanje. Kognitivno stanje se lahko definira kot zvezna spremenljivka, ali pa kot lažje določljiva diskretna spremenljvika, ki določa dva, tri, ali več vrednosti. Temu primerno se izvede tudi reševanje oziroma napovedovanje, ki je bodisi regresijski ali klasifikacijski problem. Algoritmi, ki smo jih uporabili za napovedovanje so: k-najbližjih sosedov (k-NN, angl. *k nearest neighbours*), metoda podpornih vektorjev (SVM, angl. *support vector machine*), naključni gozdovi (RF, angl. *random forest*), naivni Bayes in globoke nevronske mreže. Slednje, za razliko od ostalih, lahko delujejo na surovem signalu, torej brez izločevanja značilk. To omogoča ustrezna zgradba nevronske mreže po nivojih – v našem primeru sta primarna nivoja sledeča: konvolucijski nivo in LSTM nivo (angl. *long-short term memory*).

III Eksperiment

Cilj našega dela je ugotoviti kognitivno obremenjenost človeka z uporabo brezžičnih signalov. V okviru dela smo izpeljali študijo, kjer je triindvajset prostovoljcev individualno reševalo naloge različnih težavnosti, medtem ko je Wi-Mind sistem brezžično zajemal njihove življenjske znake. Naloge so bile prirejene iz že podobno izpeljane študije [16], ki pa je temeljila na intruzivnih merilnih metodah. Tipov nalog je bilo šest, vsaka pa je bila dana v treh težavnostnih stopnjah. Po vsaki nalogi je uporabnik izpolnil NASA-TLX vprašalnik, s katerim je ocenil svoje stanje med reševanjem naloge. Pred vsako nalogo je uporabniku predstavljen napis, pri katerem naj bi uporabnik

šel v stanje sprostitve. Ta faza naj bi določala časovno okno, ki razpolavlja časovni okni reševanja nalog pred in po pavzi - na ta način se fiziološki signali med sosednimi nalogami naj ne bi prekrivali.

Vsakemu prostovoljcu je bil na začetku predstavljen celoten potek študije, s katero so se morali tudi strinjati in privoliti v uporabo podatkov za namen naše raziskave. Med študijo je vsak na roki imel tudi Microsoft Band (pametna zapestnica), s katero smo merili srčni utrip, ki je služil kot referenca in smo ga uporabili za primerjavo utripa iz Wi-Mind sistema. Med študijo smo spremljali ekran, ki je prikazoval trenutno stanje ter v primeru težav odgovorili na vprašanja uporabnika.

Študije se je prostovoljno udeležilo triindvajset ljudi, starih od 20 do 38, sedemnajst moškega in šest ženskega spola. Povprečen čas za dokončanje študije enega prostovoljca je bil okoli 45 minut. Ključni korak je bil najti vsaj dvajset prostovoljcev, ki pa smo jih pridobili prek osebnih kontaktov.

IV Rezultati

Prva točka rezultatov je primerjava srčnega utripa pridobljenega z Wi-Mind-om in Microsoft Band-om. Primerjave kažejo različne rezultate pri različnih uporabnikih, saj razni človeški gibi prinesejo veliko šuma pri brezžičnem signalu in ga posledično ustvarijo neberljivega. Pri primerjavi se vidi trend naraščanja in padanja utripa, kar kaže na pozitivne rezultate izločanja utripa iz Wi-Mind-a. Opomniti je treba tudi, da Microsoft Band nima 100-odstotne točnosti. V našem primeru je bil za uporabo najbolj primeren Microsoft Band, zaradi njegove nizke cene, dosegljivosti in kompletom za razvoj programske opreme, ki nam je omogočil pisanje aplikacije za Android OS.

Statistike reševanja nalog kažejo na daljša reševanja nalog, ki so jih oblikovalci označili kot težje, kar pa ne drži za tip naloge *iskanje A-jev*. Podoben rezultat je viden tudi pri samooceni, ki kaže, da je samo pri temu tipu naloge najvišja težavnost bila ocenjena višje kot pa srednja težavnost. Samoocene kažejo tudi na različne razpone na lestvici, ki so jih uporabniki uporabljali

za samoevalvacijo. Zaradi tega je težko direktno primerjati in napovedovati TLX (angl. *task load index*) vrednost, lahko pa izločimo ekstremne vrednosti. Ugotovili smo, da uporabniki v povprečju dihajo počasneje med pavzo, kot pa med reševanjem naloge, kar je bilo ugotovljeno tudi v drugih študijah [5].

Za evalvacijo Wi-Mind sistema smo uporabili orodji Orange [17] in Keras [18]. Slednji se v našem primeru uporablja za učenje z nevronskimi mrežami na surovih podatkih. Želimo preizkusiti sistem za ugotavljanje kognitivne obremenjenosti človeka, zato za validacijo uporabimo *leave one person out* validacijo. Ker pa imamo podatke, ki so neenakih razponov (na primer različni TLX razponi, različne frekvence dihanja), poizkušamo tudi grupirati podobne ljudi skupaj in nato šele izvedemo evalvacijo sistema.

Osnovni problem, ki smo ga poskušali ugotoviti, je razlikovanje med stanjem sproščanja in stanjem reševanja naloge. Z metodo naključnih gozdov in naivnim Bayesom smo dobili klasifikacijsko točnost okoli 70%, z nevronskimi mrežami in surovim signalom pa 75%. Personalizirani testi pri nekaterih osebah pokažejo boljše rezultate.

Naslednji klasifikacijski problem je ločevanje med tranzicijama *v stanje sproščanja* in *iz stanja sproščanja*. Naša glavna predpostavka je ta, da ljudje začnejo dihati počasnejše oziroma hitrejše, zato smo kot značilko iz signala vključili tudi spremembo frekvence dihanja (t.j. razlika med frekvenco dihanja druge polovice časovnega okna in frekvenco dihanja prve polovice okna). Z naključnimi gozdovi smo prišli do klasifikacijske točnosti 66.4% oziroma 68% (prva vrednost predstavlja podatkovno množico brez filtriranja šumnih časovnih oken, druga pa brez omenjenih časovnih oken). Pristop z nevronskimi mrežami pride do podobne točnosti, in sicer 68.1%. Ker uporabniki dihajo z različnimi frekvencami, smo jih nato grupirali v skupini, kjer je ena predstavljala tiste, ki imajo manjšo razliko v frekvenci dihanja med stanjem sproščanja in reševanjem naloge ter v skupino, kjer so imeli večje spremembe dihanja. Pri skupini, ki je imela večje spremembe dihanja smo prišli do večje klasifikacijske točnosti (69.5%), saj algoritem lažje zazna spremembe v dihanju.

Naslednji izziv je ugotavljanje težavnosti naloge iz brezžično zajetih signalov. Brez razdelitve celotne podatkovne množice, kjer imamo več tipov nalog in težavnosti, je težko napovedati trenutno težavnost naloge. Do pozitivnejših rezultatov pridemo, če naredimo modele po posameznih tipih nalog in, če odstranimo *srednjo* težavnost nalog, ki se včasih očitno prekriva s sosednjima težavnostnima stopnjama. Pri tipu naloge *primerjanje številk* lahko z naključnimi gozdovi s 65.4% klasifikacijsko točnostjo napovedamo težavnost naloge (*lahka* ali *težka*), kar je za dobrih 15% boljše od naključnega klasifikatorja. Značilka, ki najbolj prispeva k temu rezultatu, je HRV visokih frekvenc (izračunano po *information gain* vrednosti).

Z regresijskim problemom za napoved TLX vrednosti, se zelo težko približamo pozitivnim rezultatom. Najboljši rezultat dobimo pri *testu sledenja črti*, kjer pa najbolj vplivajo značilke povezane z dihanjem.

V Diskusija in zaključek

Cilj dela je bil ugotoviti kognitivno obremenjenost človeka z uporabo brezžičnih signalov. Izpeljali smo študijo na triindvajsetih prostovoljcih, ki so reševali naloge različnih težavnosti, medtem pa smo brezžično zajemali njihove življenjske znake (t.j. dihanje in srčni utrip). Iz brezžičnih signalov smo nato poskušali izločiti srčni utrip, vendar nismo bili pri vseh uporabnikih tako natančni kot so bili avtorji Vital-Radia [9]. Eden od razlogov je ta, da se pri premiku rok signal zelo popači in ga ustvari neberljivega. Zanima stvar je ta, da smo s cenejšo opremo, v primerjavi z avtorji Vital-Radia, lahko z razdalje izločali dihanje in bitje srca. Slednje pa je bilo tudi temu primerno, vendar še zmeraj primerljivo.

Izboljšave evalvacije sistema bi lahko vključevale višje število udeležencev študije, saj večji vzorec predstavlja bolj reprezentativne rezultate. Različne težavnosti so bile pri posameznih tipih nalog premalo jasne, zato bi ena od izboljšav lahko vključevala le dve težavnosti, ki bi imeli večji težavnostni razpon. Nekatere naloge so bile s strani uporabnikov rešene zelo hitro (na

primer manj kot pol minute), kar postane težko za povezavo teh časovnih oken s kognitivno obremenjenostjo. Ena od možnih izboljšav bi lahko imela naloge, ki se jih rešuje dalj časa (na primer pet minut), ker se HRV frekvenčne značilke izrazijo oziroma so bolj razvidne na daljših časovnih oknih.

Chapter 1

Introduction

As our reliance on ubiquitous computing devices grows, so does the need for seamless interaction with these devices. The postulates defined by Mark Weiser in 1991 call for "calm" technology that blends in with the environment, understands the user, and works towards fulfilling the user's needs [19]. Unfortunately, almost thirty years later we are surrounded by a plethora of devices that remain completely oblivious to our needs, and that contradict Weiser's vision by getting in the way of our actual intents. Mobile communication devices are a prime example of such a conflicting technology, as an average smartphone user receives around 100 push notifications per day, most of which are disruptive [20]. Through these notifications and other disturbing signals users attention is fragmented, which reduces work performance [1] or induces stress [2]. Bringing experiences from the field of cognitive perception closer to ubiquitous computing developers is a difficult task and we have to be aware that machines do not understand us or have difficulties to do so.

Understanding a human user encompasses multiple aspects of human consciousness, from sensing one's emotions, over inferring one's goals, to perceiving one's fatigue. Recent research, however, has shown the link between a user's interruptibility and her immersion in a task at hand [21, 22], making the inference of mental effort a promising potential enabler of improved human-computer interaction (HCI). Attention management system might

control users attention in case of interruptions by choosing suitable environment change or postponing interrupting notifications to a later time [23]. The core of these systems are sensors (to acquire users current physiological signs), machine learning algorithms (to learn users interruptibility or mental effort) and actuators (to make suitable action regarding users cognitive load level).

Inferring someones cognitive load is challenging and can be done in multiple ways, e.g. by subjective self-evaluation after completing some task or by observing the person's performance on the task. One example for such measurement is NASA-TLX (Task Load Index), where participants report their load after completing a task [3]. However, these highly subjective evaluations can be also correlated with more objective physiological signals, which are results of a human autonomic nervous system and cardiovascular system reaction. Some of these signals include heart rate blood pressure [24], heart rate variability [4], respiratory changes [5], brain activity [6], galvanic skin response (GSR) [25, 26], eye movement [6], pupil size, and facial expression [27]. These can be measured with special equipment, e.g. nasal thermistor, chest respiration strap, ECG (Electrocardiogram), sphygmomanometer (blood pressure monitor), smart watch, electroencephalography (EEG), etc. One thing in common for all these monitors is – they are intrusive, i.e. they require a body contact.

While to date research in understanding one's mental effort has been tested mostly on intrusive methods, with notable exceptions of camera-based approaches [7, 8], here we explore the prospects of devising a *wireless* non-intrusive vital sign radar monitor to infer a user's cognitive load. We design and implement a software-define radio-based wireless system prototype and through real-world experiments on a group of 23 volunteers evaluate its ability to sense physiological signals and through machine learning connect these to a user's mental effort. The contributions of our work are the following:

- we identify and adapt radar technology for the purpose of vital sign monitoring using software-defined radio concept;

- we extract vital signs (related to breathing and heart activity) from the raw signal and evaluate heartbeat detection accuracy;
- we conduct a user study to collect vital signs using our system while the users are solving tasks of different types and complexities;
- we use machine learning algorithms to determine users cognitive load engagement with the acquired wireless signal data.

Our system for wireless cognitive inference is called Wi-Mind and includes a software-defined radio-based frequency modulated carrier wave radar, data processing and feature extraction, and a machine learning pipeline.

This thesis is structured as follows. In Chapter 2 we present related work on the topic of cognitive load inference using non-intrusive methods. In Chapter 3 we describe our proposed system for wireless cognitive load inference, where some of the main concepts, such as radar type, used hardware, software and wireless signal processing are described in detail. In Chapter 4 we describe our experiment approach, where we used our system to perform the study. Chapter 5 shows heartbeat benchmark, descriptive statistics about user study and machine learning accuracies using different algorithms to determine different cognitive load engagement. In Chapter 6 we discuss our approach and limitations and finally, in Chapter 7 we conclude our work.

Preliminary results of Wi-Mind system were presented at the "3rd International Workshop on Smart & Ambient Notification and Attention Management" [28]. In this thesis we describe this system and the whole study in detail and introduce more approaches to evaluate the whole system.

Chapter 2

Related work

HCI technology is focused on the interfaces between people and computers and enables improved interaction with computing devices. This interaction can be thought of as a dialogue; computer and people alike can handle only a limited amount of information. Exceed the ability to handle huge amounts of information can lead to worse task performance, miss important data or even abandoning some tasks [29]. Human cognitive processing capabilities are limited by our physical resources. These resources include, e.g. visual, aural, motor resources, but also procedural and declarative memory resources. Even simple tasks require multiple kinds of resources, and many tasks can bring complex patterns of interferences between such resources [23]. Resources are independent and can be accessed in parallel, yet, are exclusive, and a single resource can be assigned to a single task at a time [30].

While a task competes for resources, interruptions arise when a stimulus signals a new task. **Interruptions** can be considered as new tasks (i.e. secondary tasks) on top of the main current activity (also called the **primary task**), which results in loss of attention in the current task [31]. The primary task has to be stored in declarative memory (i.e. one of the long-term human memory types which is used to recollect previous experiences and concepts [32]) and will be retrieved after the interruption is handled. The complexity of the interrupted task influences the perceived disruption and

the ability to successfully retrieve the task. This can be reflected in delayed task completion and task errors [30, 33]. One way to quantify the task complexity is through cognitive load measurements. Term **cognitive load** is by Paas and van Merriënboer defined as "*a multidimensional construct representing the load that performing a particular task imposes on the learner's cognitive system*" [34].

In cognitive load research users' performance on the secondary task is often used as a measure of cognitive load elicited by the primary task [23, 34]. Cognitive load can be assessed by measuring mental load (interaction between task and subject characteristics), mental effort (allocated capacity to accommodate the demands imposed by the task) and performance (users achievements) [34]. Some of the cognitive load measurement methods gather data on the subjective perception of task difficulty, performance data using primary and secondary task techniques and psychophysiological data [34]. Measuring subjective data is performed using surveys (e.g. NASA-TLX [3] and SWAT – Subjective Workload Assessment Technique [35]) which are solved by user at the end of a task. While subjective rating scales are relatively unintrusive, they cannot be used in real time. Physiological techniques measure physiological variables, such as heart rate variability, brain activity and eye activity. Measuring interruptibility using biometric sensors, such as headband and wristband, has been conducted by Zücker et al. [36]. They discovered that EEG signals, eye blinks, skin conductance, heart rate and inter-beat interval features show positive correlation between interruptibility and mental load. Gjoreski et al. conducted a study to detect stress, using commercial wristband and extracting various heart rate features [37]. Their findings show that the approach is quite reliable on a two-class problem (i.e. stress versus no stress class), however, it has still some room for improvement. The authors of the mentioned articles show that cognitive load correlates with changes in vital signs. However, vital signs monitoring in those studies were made with intrusive methods.

Recent advancements in technology enabled non-intrusive vital signs'

monitoring, such as camera-based approaches to measuring heart rate variability (HRV) [8] and detecting pulse from head motions in a video [38]. In 2015, Adib et al. introduced Vital-Radio [9], a wireless sensing technology for monitoring breathing and heart rate without body contact that exploits the fact that wireless signals are affected by the motion in the environment. More specifically, chest movements due to human inhaling/exhaling and skin vibrations due to heartbeats (a process of inferring heartbeats using chest movement is called ballistocardiography – BCG) can be captured by observing reflected radio waves' phase variation. Similar wireless-based vital signs monitoring systems include TensorBeat [10], which employs channel state information (CSI) phase difference data to estimate breathing rates for multiple persons with commodity WiFi devices, WiBreathe by Ravichandran et al. [11], an ultra-wideband (UWB) radar by Huang et al.[12], and impulse-radio (IR) UWB Doppler radar-based solutions [13, 14]. Another similar approach was introduced by Nandakumar et al., where they transformed smartphone into an active sound signal emitter/listener to detect sleep apnea [15].

In terms of applications, Zhao et al. used a technology similar to Vital-Radio, called EQ-Radio, for analysing radio frequency (RF) reflections off a person's body to recognize the emotional state [39]. To infer cognitive load unobtrusively, Abdelrahman et al. use thermal imaging cameras focused on a persons forehead and nose [7], while McDuff et al. use a five-band digital camera to detect cognitive stress [8]. The latter is using the concept that cognitive tasks have an impact on breathing and HRV. While promising, the need for frontal camera placement might limit the applicability of the above approaches (e.g. for inferring a car driver's engagement). Urh and Pejović use smartphone sensing to infer task engagement, however, their work remains at a coarser granularity as it, among other features, concentrates on location, time, and calendar events [40].

As mentioned, experiments with measuring vital signs, such as heart rate and heart rate variability, have shown that they correlate with users work-

load. However, measuring vital signs with wireless signals to assess users cognitive load, has not, at least to our knowledge, been attempted before. The goal of our work is to explore and implement a radar-based monitor to infer cognitive load wirelessly.

Chapter 3

Wi-Mind System

We implemented Wi-Mind system, which is based on the premise that a person's vital signs, such as respiratory rate and heart rate, correlate with that person's cognitive load. The system consists of two main modules: *wireless monitoring module* and *machine learning module/feature extraction module* (see Figure 3.1). *Wireless monitoring module* collects raw vital signs data and *machine learning module* extracts and infers one's cognitive load based on the collected data. A user is stationary (seated) in an office setting and engaged in a mental task. One antenna of the wireless module is placed on the right, the other on the left side of the person, and are used to unobtrusively obtain data corresponding to the users movement, which in turn conveys into vital signs. The data is further filtered and processed, and forwarded to the machine learning module that then makes the final inference about the person's cognitive load.

In the following sections, we describe the methods, hardware and software used in this work. While focusing on the ideas behind each method, we also try to describe our approach from a practical point of view. The experiment setup and study details is described in Chapter 4. Wireless vital sign monitoring accuracy and results are discussed in Chapter 5. The code, that was used to implement and evaluate our system is available on our Github repository [41].

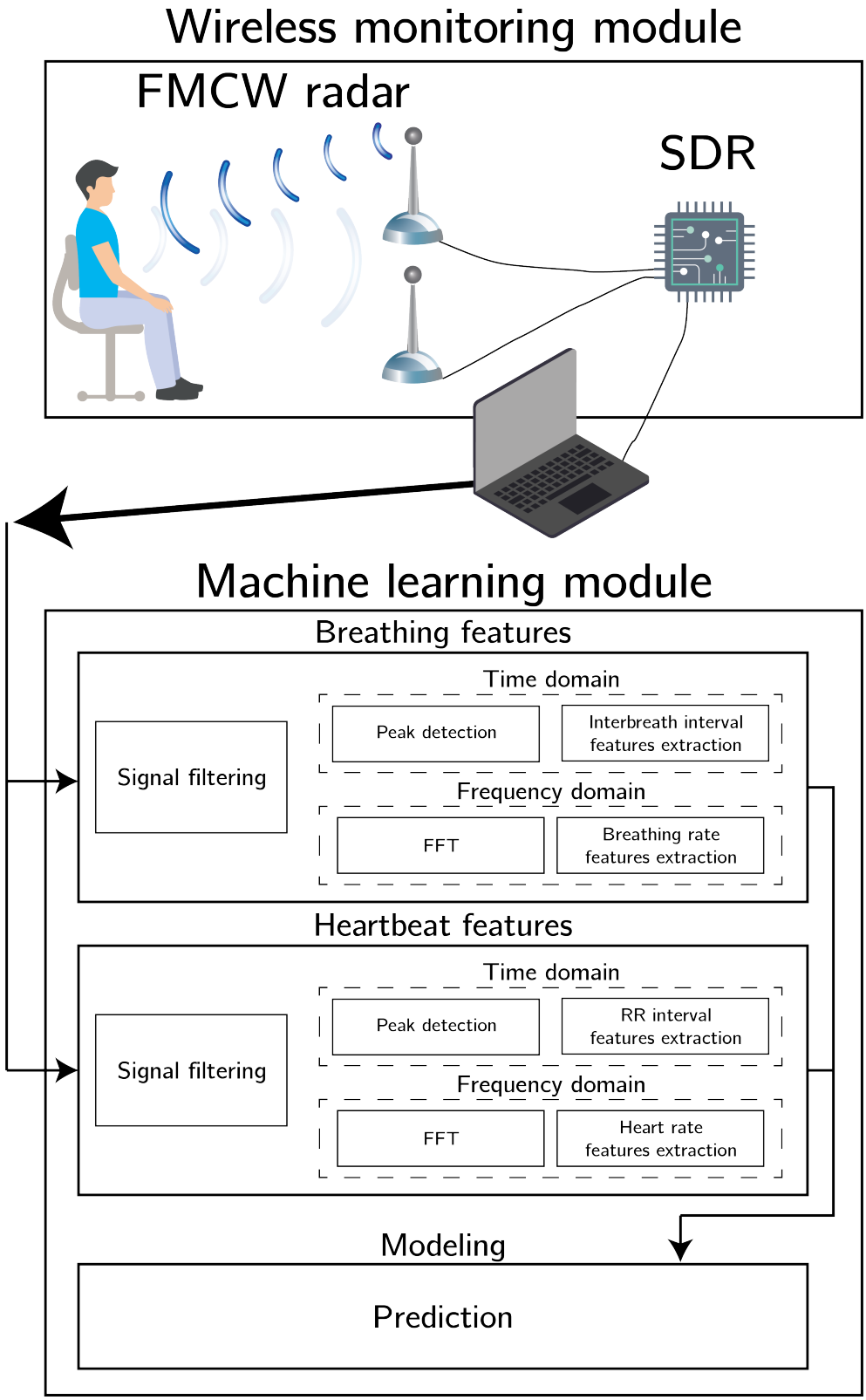

Figure 3.1: Wi-Mind scheme - wireless cognitive load inference system that relies on a software-defined radio-based FMCW radar and a machine learning data-processing pipeline.

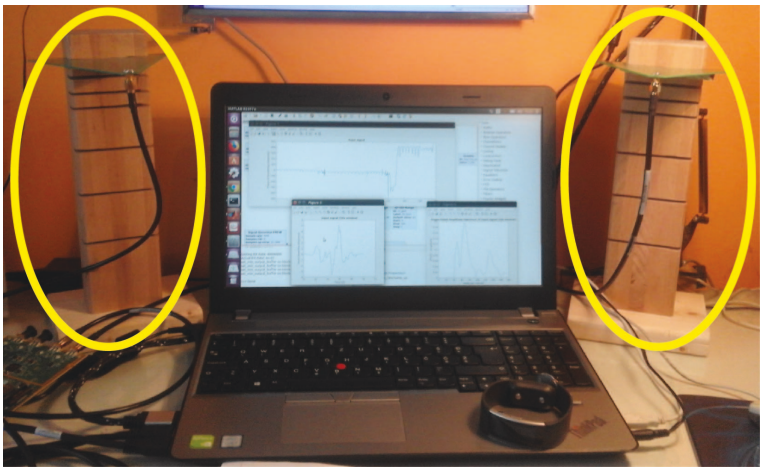

Figure 3.2: Wi-Mind system and its antennas, marked with yellow ellipses.

3.1 Wireless monitoring module

The idea for unobtrusive vital signs data collecting was taken from the already mentioned Vital-Radio system [9]. Recent advancements in central processing unit (CPU) capabilities and signal processing algorithms have led to *software defined radio* (SDR) – a concept that enables highly customizable transmission/reception through a symbiosis of radio front-end hardware and signal processing on a general purpose computer. The core of Wi-Mind is an SDR implementation of a radar that is based on a slightly modified *gr-radar* [42] FMCW module running on top of the GNU Radio SDR framework [43]. Radar allows us to filter out large multipath interference and then perform a fine-grain movement analysis of the user's body (predominantly chest), which may correspond to breathing and heartbeats. The phenomenon that allows this system to detect heart rate from signal reflections is called ballistocardiography (BCG), which can represent repetitive motions of ejecting blood into vessels caused by heart. Figure 3.2 depicts hardware, used by Wi-Mind system.

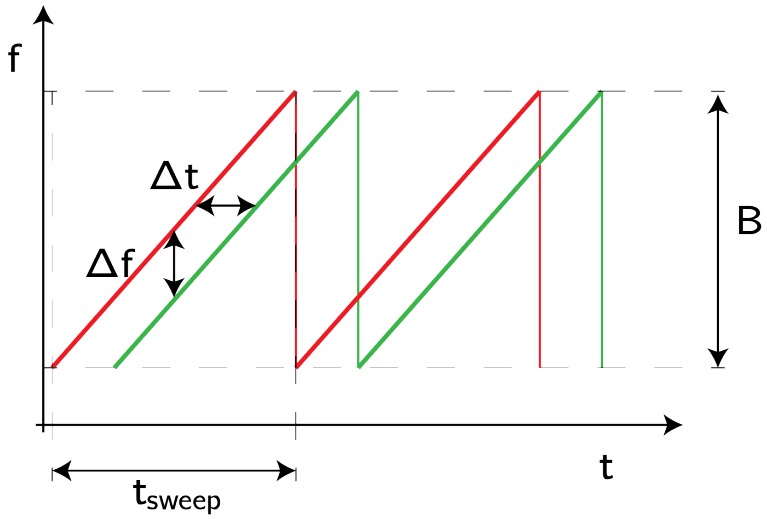

Figure 3.3: Frequency modulated carrier wave ranging example. The red line presents transmitted signal and the green line the received and delayed echo signal.

3.1.1 FMCW radar

Radar is a device capable of determining distance. FMCW radar is a type of continuous-wave (CW) radar operated on a frequency-modulated waves. It changes its operating frequency periodically during the measurement and has some advantages versus similar CW radars by providing increased reliability for distance along with speed measurement. In the FMCW method, a signal is transmitted and a delayed echo signal is received, which translates into frequency shift in comparison to the currently transmitted wave (see Figure 3.3). The difference in time between the transmitted and received signal can be estimated due to constant frequency-change slope and can tell us distance to the object, from which the signal was reflected.

FMCW transmits a narrowband signal, whose frequency changes linearly in time [44]. The frequency can either change with sawtooth (see Figure 3.3) or triangular modulation. The distance to the reflecting object can be deter-

mined with the following equation:

$$R = \frac{c \times \Delta t}{2} = \frac{c \times \Delta f}{2 \times \frac{df}{dt}} \tag{3.1}$$

where R stands for distance, c for speed of light, Δt for time delay between transmitted and reflected signals (s), Δf frequency difference, also called beat frequency (Hz) and $\frac{df}{dt}$ frequency shift per unit of time (or slope of the frequency modulation, which can also be presented as $\frac{df}{dt} = B/t_{sweep}$, where B is sweep bandwidth and t_{sweep} sweep time). From Equation 3.1 we can also distinguish distance resolution and see that it depends on the sweep bandwidth [44]. Combining Equation 3.1, $\frac{df}{dt} = B/t_{sweep}$ and $\Delta f = 1/t_{sweep}$ (frequency resolution of the FFT) gives us distance resolution:

$$\Delta R = \frac{c}{2 \times B} \tag{3.2}$$

Resolution of a radar is defined as a minimum distance of two targets that can be detected separately from the resolution of Fourier transform. From the Equation 3.2 it is obvious that we would need a very large bandwidth to detect small movements, say those corresponding to human heartbeat and breathing. Having an SDR sweep over such a wide bandwidth is not feasible with the current state-of-the-art commodity equipment. While custom solutions have been prototyped for increasing the sweeping bandwidth and narrowing the resolution to a few centimeters [9], in our work we rely on FMCW simply to filter out large multipath interference, and utilize further signal processing to extract heartbeat and breathing.

Phase extraction and analysis

Locking the signal to the specific distance (i.e. take only one bin from Fourier transform) and then taking the phase of the reflected signal, by analysing phase variation, very small movements can be resolved. The phase of the reflected signal is related to the distance traveled [9]:

$$\Phi(t) = 2 \times \pi \times \frac{d(t)}{\lambda} \tag{3.3}$$

where λ is the wavelength and $d(t)$ is the travelled distance.

Having center sweep frequency at 5.2 GHz gives wavelength of 58 mm (from equation $\lambda = c/f$, where λ is wavelength). Theoretical minimum detectable distance change with this frequency is 161 μm (according to Equation 3.3 and assuming that 1° change in phase can be detected). These subcentimeter variations in distance to the chest, which are caused by breathing and heart beats should be enough to analyse periodic jitter in the wireless signal [44]. Phenomenon, which allows us to detect heart beats, is called ballistocardiography.

Theoretical minimum distance between different objects, in order to be separated with FMCW radar, has to be larger than distance resolution (see Equation 3.2). To acquire and analyse signal phase, Fourier transform has to be calculated, where each FFT bin corresponds to different range. This brings us to one drawback of analyzing signal phase – objects that fall in the same FFT bin cannot be distinctly separated with signal phase [9, 45]. Thus, in our work we use directional antennas that are focused to a singled out seated individual. In the future, we plan to adapt the system to be able to infer cognitive load even in presence of other people.

3.1.2 SDR

Software-defined radio (SDR) is a concept for a programmable radio communication system and it usually consists of a personal computer with analog-to-digital converter and RF front-end (see Figure 3.4). This functionality produces general-purpose processor that introduces flexibility for signal processing and makes special-purpose hardware somehow obsolete. The concept is not new, however recent evolving capabilities have led to its popularity.

One of the most useful SDR advantages is its flexibility. Legacy radios are constrained with RF front-end and, unlike SDRs, do not have the capabilities to be arbitrarily programmed. If we look at smartphones and similar devices, they currently have many different radios optimised for signals operating on different frequency bands (WiFi, LTE, GSM, UMTS, GPS) [47]. Implemen-

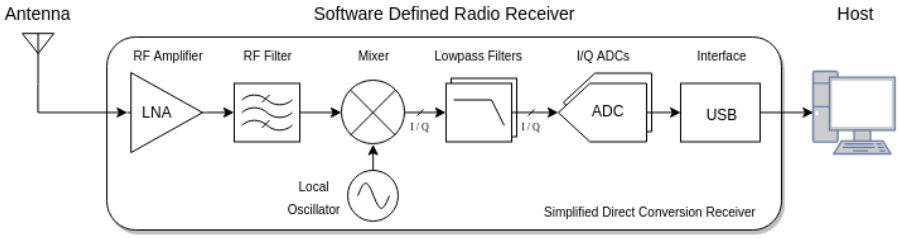

Figure 3.4: An example scheme of SDR receiver (from [46]).

tation of a single SDR in such devices could dynamically switch between various frequency bands with a software code. Some of the most popular commercially available SDR devices are USRP, HackRF and RTL-SDR.

Motivation

Because of its flexibility and our experimental nature of problem we decided to use SDR. One of the reasons for SDR is also FMCW radar. SDR allows us to do advanced signal processing that requires access to low-level radio data.

In this work, we used an SDR front end called Ettus Research Universal Software Radio Peripheral (USRP) B210. The product has RF coverage from 70 MHz to 6 GHz and is able to use multiple antennas. The nature of FMCW radar requires full-duplex transmission (radar receives the signal at the same time as it is sending signal with another frequency) – this is possible with the usage of multiple antennas. It is more reasonable in our case to use directional antennas instead of omnidirectional antennas, as they radiate and receive in specific direction, resulting in increased performance and reduced interferences from unwanted sources. One of the most common antenna type, which was also used for our system, is log-periodic antenna, allowing us to operate over a wide band of frequencies. We used LP0965 log periodic antennas.

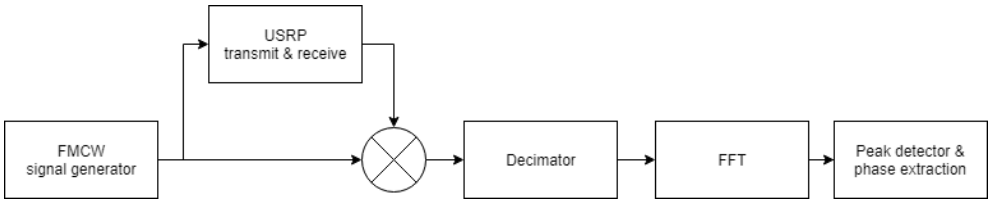

Figure 3.5: Simplified GNU Radio flowgraph for FMCW radar and phase extractor.

3.1.3 GNU Radio

GNU Radio is a free and opensource development toolkit for SDR programming [43]. Its applications are usually known as "flowgraphs" or connected blocks. It comes with already pre-built blocks for signal processing, however new custom blocks can be written either in Python or C++ programming language. GNU Radio can either be used as a simulation environment or as a real-time processing software with the suitable SDR hardware.

Implementation

As already mentioned, the most promising method for vital sign monitoring in our case was radar that merges the ideas from FMCW and phase analysis (see Section 3.1.1). The core GNU Radio implementation of FMCW radar was taken from the existing *gr-radar* module [42] (see Figure 3.5 for a simplified GNU Radio flowgraph). Our main radar setup values are:

- center sweep frequency: 5.2 GHz
- sweep frequency bandwidth: 20 MHz
- frequency modulation pattern: sawtooth
- sampling frequency: 40 MHz
- decimation factor: 8 (decimator is used to reduce computational complexity)

Combining sweep frequency bandwidth and Equation 3.2 gives us distance resolution of 7.5 meters (i.e. minimum distance of two targets that can be detected separately) and center frequency returns sub-centimeter variations in chest distance due to breathing, which are caused by sub-radian variation in the phase [9]. Our distance resolution (7.5 meters) is in comparison with Vital-Radio [9] resolution way smaller (8 centimetres). Our implementation is not able to differentiate between multiple objects in e.g. range of 7.5 meters from the antennas, but is enough to filter out large multipath interferences and measure vital signs of one user, since all other reflections in the room should be static (i.e. are the result of walls or other static objects). Originally, the *gr-radar*'s FMCW radar was implemented to calculate the actual distance of the reflecting object (in meters). Minor changes in the original *gr-radar* implementation were made:

- we locked the distance to specific range (i.e. FFT bin);
- we extracted only the phase of the signal and timestamps for each phase sample.

The whole GNU Radio implementation (counting sampling frequency, decimation factor, bandwidth, and other factors) gave us 43 samples per second in the raw signal, which should be enough to analyse breathing and heart rate.

3.2 Machine learning module

We are trying to predict users' cognitive load based on their vital signs, acquired with wireless radar monitor. Since we are dealing with raw signals reflected of a human body (with radar technique described in Section 3.1.1), we have to extract useful features out of them. In this section we describe how we filtered the raw signal, extracted breathing and heartbeat features and used this data in machine learning algorithms to estimate mental effort. See Figure 3.6 for a preview of this module.

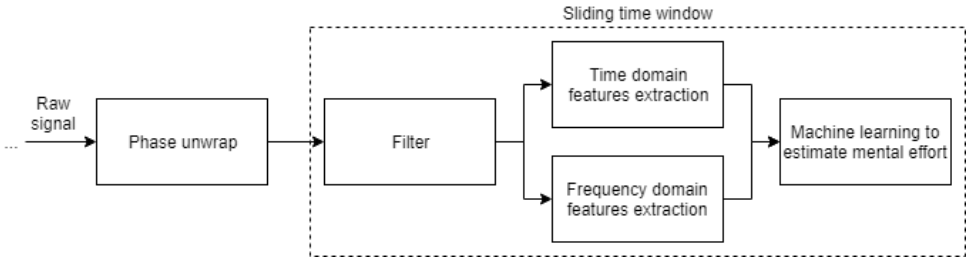

Figure 3.6: Features extraction and machine learning module.

3.2.1 Data preprocessing

In our experimental Wi-Mind system, GNU Radio program saves the raw signal phase shift and its sample timestamps into two separate files.

The reflected signal phase contains information about a users breathing and heartbeats as the phase corresponds to the distance the wave has travelled. But the phase in the output file is constrained to the interval $(-\pi, \pi)$ yielding so called *wrapped phase*. That means that when the next adjacent value exceeds π, it continues on the other side $(-\pi)$ and vice versa. Since this results in switching signs and makes the breathing signal unrecognisable, we have to make a continuous function out of this signal and this is called *unwrapped phase*. The idea for phase unwrapping can be seen in Figure 3.7.

Now we can deal with breaking the signal into time frames and proceed with filtering to extract breathing and heartbeat features.

3.2.2 Feature engineering

This section describes our approach towards extracting relevant features out of raw signal. Since human breathing and heartbeat have different features, we described them separately.

Breathing features

Raw signal contains low frequency and/or high frequency noise. The former can be caused by very slow (slower than average breathing rate) body

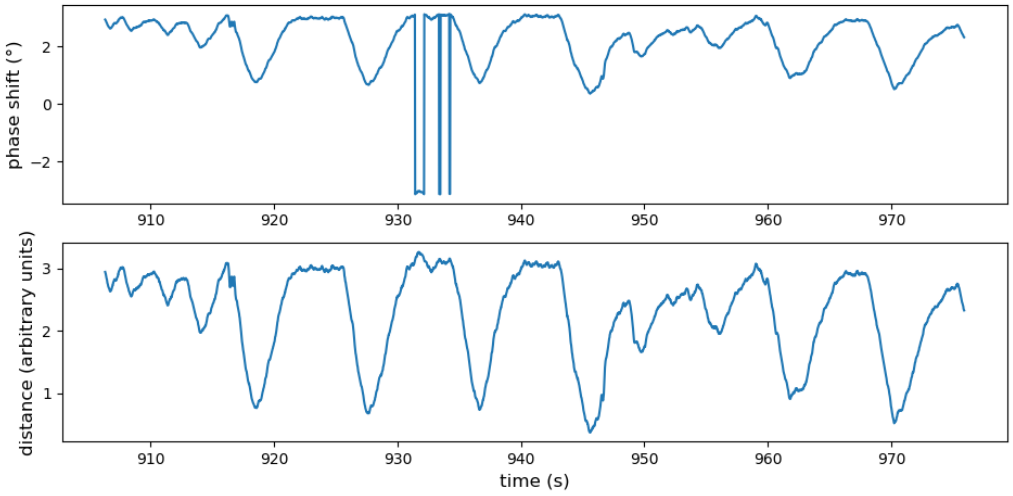

Figure 3.7: Wrapped phase (top) and unwrapped signal phase (bottom).

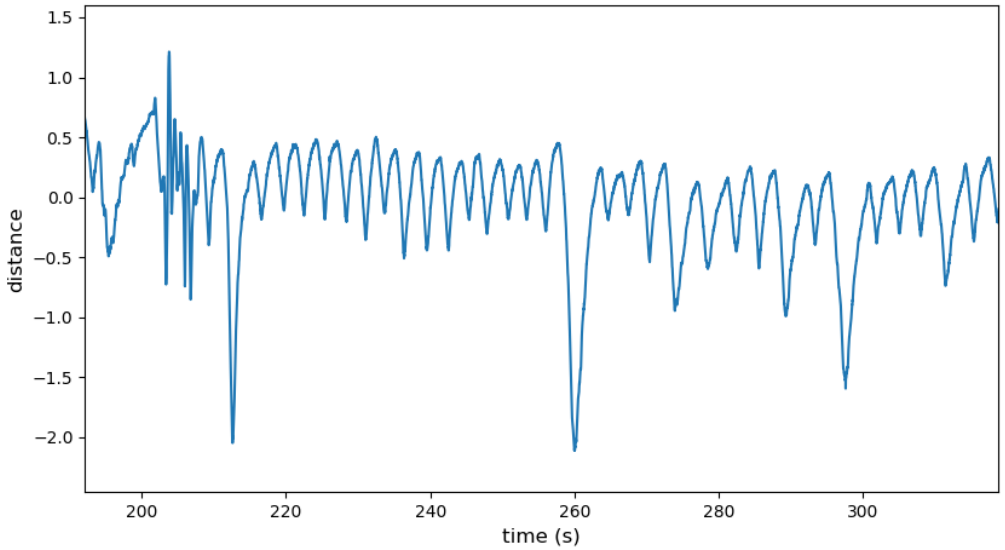

Figure 3.8: High frequency noise at around 205 seconds and slow signal drift through the whole time.

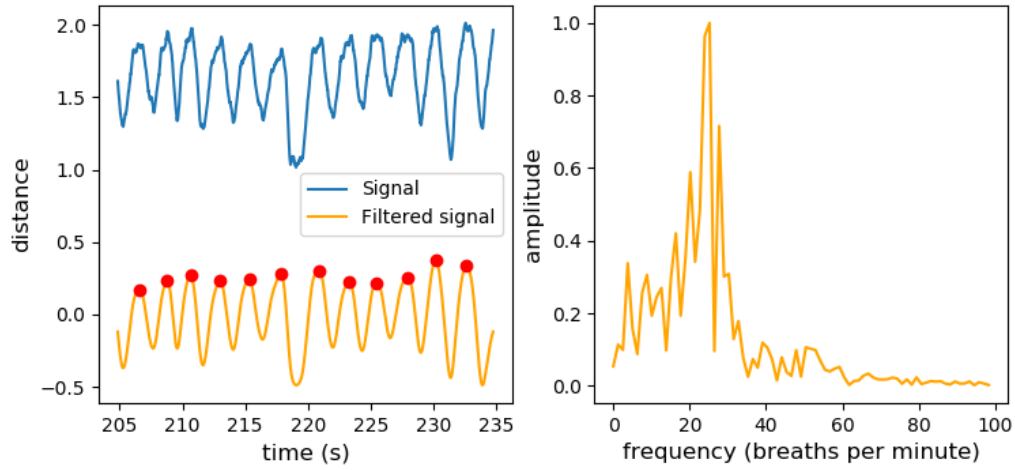

Figure 3.9: An example of signal filtering (left) and its frequency domain (right) on a 30 second time window, where maximum peak corresponds to the current breathing rate. Each red dot presents one breath.

movements and the latter with very fast movements, limb motion or other random noise (see Figure 3.8). To cope with this noise we use filters. To remove slow signal drift and high frequency noise, we use a band-pass filter – anything below 0.083 Hz (5 breaths per minute) and above 1 Hz (60 breaths per minute) is eliminated (see Figure 3.9). This frequency range was chosen because the average breathing rate of an adult human is around 12 to 20 breaths per minute and the filtered signal should "erase" non breathing noise.

The most straightforward respiratory feature is the breathing rate. To extract the breathing rate from a specific time window we calculate the Fast Fourier Transform (FFT) of the signal and then single out the highest peak in the frequency domain (see Figure 3.9). The position of the peak corresponds to the breathing rate – i.e. if a person's respiratory rate is 20 breaths per minute, then the FFT will have the highest peak at the value 20.

Next frequency domain feature would be the difference between the average breathing rate at the first half and the average breathing rate at the

second half of a time window. This could indicate the start or the end of solving a mental task as people might speed up or slow down their breathing rate. We can also look at spectral power features from the respiration signal representing the energy in the following bands of: 0.1-0.2 Hz, 0.2-0.3 Hz, 0.3-0.4 Hz and 0.4-0.5 Hz. Some of these bands were found useful in [48]. With the calculated energy in the mentioned bands we could determine if somebody is breathing with different breathing rates (i.e. if one is, in one time window, breathing slow at first and then speeds up breathing rate, then this would show in different band energies).

Time domain features can yield additional information. One possibility is to get respiratory rate variability (RRV). The idea is to measure inter-breath interval, i.e. intervals between each inhale or exhale. To detect times at which inhales happen, we use a peak detector on a filtered signal (see Figure 3.9). From marked timestamps we then calculate time differences between breaths. One of possible extracted features for breathing is the average of these intervals, but this highly correlates with the average breathing rate, calculated from the frequency domain. Standard deviation of these intervals could tell us if individual is breathing equally or is sometimes holding breath (medical term for involuntary holding breath or cessation of breathing is called apnea).

Calculating some statistical measures from raw signal (not looking at the frequency domain or calculating breath peaks) has also been considered: mean, median, standard deviation and root mean square value (as seen in [49, 50, 51]).

To deal with noise (e.g. limb motion) we introduce a meta-feature, which is set to *true* if the time window does not contain too much noise and *false* vice-versa. To determine the noise, we choose the FFT peak frequency (see Figure 3.9 right) and calculate if the peak value is at least five times higher than the average power in the remaining frequencies (as seen in [9]).

All of the respiration related features can be seen in Table 3.1. Our implementation makes a new estimate, on a 30 second long time window, each second. To additionally cope with noise we implemented the rolling

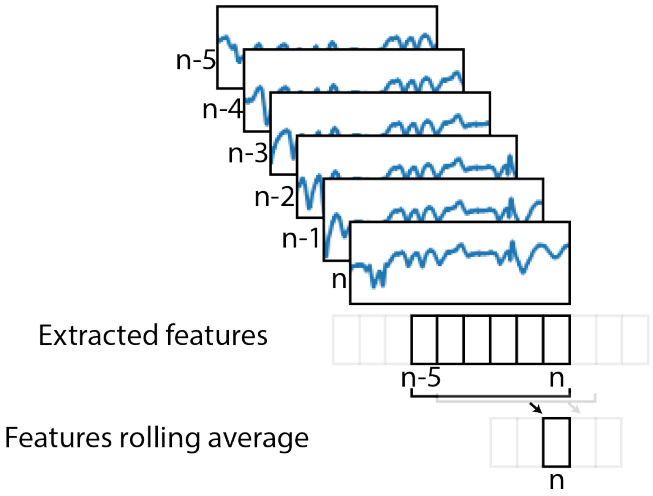

Figure 3.10: A sliding time window approach to signal feature extraction and averaging.

average of all features for last six sliding time windows (this number was chosen empirically) – see Figure 3.10 for a visual idea of this process.

Heartbeat features

Similar to breathing features extraction, we again filter the raw signal with a band-pass filter to extract heartbeat features, but this time with the cut-off frequencies of 0.83 Hz and 2.5 Hz (50 to 150 beats per minute as an average adults heart rate is in range of 60 to 100 beats per minute) – see Figure 3.11.

The average heart rate in a given time frame can again be calculated by extracting the highest peak in the frequency domain (or the second highest peak using filter to extract [40-200] beats per minute, as noted in [9], since the highest peak is due to the leakage from the breathing) – see Figure 3.11. Difference in heart rates between the beginning and the end of a time window has also been considered. Research has shown that heart rate increases during stressful times [51, 52]. However, these metrics can not be directly comparable as people can have different heart rates based on their current

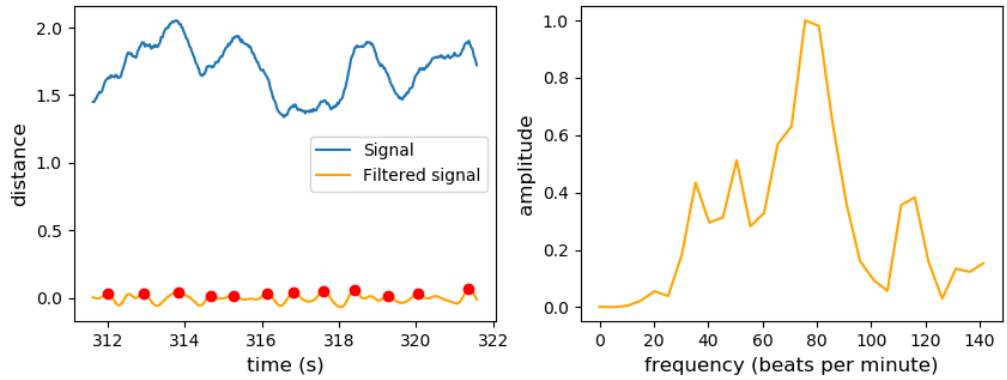

Figure 3.11: An example of signal filtering (left) and its frequency domain (right) for heartbeat extraction on a 10 second time window. Each red dot presents one heart beat and the maximum value in the frequency domain graph presents the most probable heart rate.

state, fitness or physical condition. Somewhat more useful would be the heart rate variability (HRV).

With HRV we have in mind the variability of time intervals between successive beats (also called RR intervals or NN intervals) and this can relate to emotional arousal, strain, attention and motor inhibition [53]. HRV can either be low (constant intervals between heartbeats) or high (interval lengths variate). High HRV is correlated with relaxed situations and low HRV is correlated with stress situations. This is a result of the autonomic nervous system, which unconsciously regulates body functions, such as heart rate, digestion, respiratory rate etc. Timestamp for each heartbeat can be marked a with peak detector (see Figure 3.11) and NN intervals can be calculated as time differences between adjacent timestamps. To deal with outliers, we removed all NN intervals that differed more than 25% from the preceding NN intervals (as seen in [54]). Some of time domain measures, calculated from NN intervals are:

- mean NN (mean value of the NN intervals, as seen in [55]);

- SDNN (standard deviation of NN intervals);
- RMSSD (root mean square of successive differences)
- pNNx (proportion of the number of pairs of successive NNs that differ by more than x ms; x = 50, 70).

HRV (variations between sequences of consecutive heart beats) can also be calculated in frequency domain. Instead of using classical FFT, a more appropriate frequency estimation method is the Lomb-Scargle (LS) periodogram. Although it is a slower than FFT, LS can produce more accurate estimates for typical NN data, as it can work with unevenly sampled data. Some of the typical frequency domain HRV features are ([55]):

- low frequency (LF) - 0.04-0.15 Hz;
- high frequency (HF) - 0.15-0.4 Hz;
- LF/HF ratio.

All of the heartbeat features can be seen in Table 3.1. Each second a new estimate over a 10 seconds long time window is made for most of the features. HRV frequency domain features require longer time window (100 seconds in our case). However, extracting heartbeat features is less reliable than extracting breathing features, as heart beats are harder to detect, but we still try to extract some additional information from the signal. To deal with noise, we introduce a rolling average to all features for the last 10 sliding time windows (the number was chosen empirically).

3.2.3 Modelling

Our goal is to predict a user's cognitive load, which is in this work assessed with the task load index metric (a continuous variable). However, as we are tackling an extremely challenging problem with rather experimental equipment, we need to examine the ability to solve a more coarse grain classification problem, such as classification between busy/relax time frames and

Table 3.1: All of the breathing and heartbeat features extracted from signal and then later used in machine learning modelling.

breathing features	feature label	meaning
breathing rate	br_rate	mean respiration frequency
breathing rate difference	br_rate_difference	change in respiration frequencies from first half of time window and second half
spectral power in bands of		area in normalized frequency domain between a range of frequencies
0.1-0.2 Hz	br_freq_6_12	
0.2-0.3 Hz	br_freq_12_18	
0.3-0.4 Hz	br_freq_18_24	
0.4-0.5 Hz	br_freq_24_30	
mean inter-breath interval	br_IBI_mean	mean value of inter-breath intervals
STD of inter-breath interval	br_IBI_std	standard deviation of inter-breath intervals
mean of raw signal	br_raw_mean	mean value of raw signal
median of raw signal	br_raw_median	median value of raw signal
STD of raw signal	br_raw_std	standard deviation of raw signal
RMS of raw signal	br_raw_rms	root mean square value of raw signal
noise filter	br_ok	meta feature to determine if the time window is clean or noisy
heartbeat features		
heart rate	hr_rate	mean heart rate
heart rate difference	hr_rate_difference	change in heart rates from first half of time window and second half
mean NN	hr_NN_mean	mean NN interval value
SDNN	hr_SDNN	standard deviation of the adjacent NN intervals
RMSSD	hr_RMSSD	the square root of the mean of the squares of the successive differences between adjacent NNs
pNN50	hr_pNN50	the proportion of NN50 (number of pairs of successive NNs that differ by more than 50 ms)
pNN70	hr_pNN70	the proportion of NN70 (number of pairs of successive NNs that differ by more than 70 ms)
HRV frequency features		
LF	hr_HRV_lf	heart rate variability in the 0.04–0.15 Hz band
HF	hr_HRV_hf	heart rate variability in the 0.15–0.4 Hz band
LF/HF	hr_HRV_lf_hf	ratio of the low and high frequency of heart rate variability

transitions between them. In classification problems we attempt to determine the final class, which is in our case a variable of two or three categories, from values of independent features (e.g. breathing and heartbeat features, see Table 3.1). In regression analysis we try to estimate dependent variable, which is a continuous function (e.g. cognitive load estimation). The results and the validation accuracies were made offline, i.e. we collected data from users (see Chapter 4) and then we evaluated the results in Chapter 5.

Classification and regression algorithms

Here we list some of the machine learning algorithms to evaluate Wi-Mind system.

- **k-NN** (k nearest neighbours)

  k-NN is a type of lazy learning algorithm, where the final prediction is based on the k closest training examples in the feature space. It has a k parameter, which has to be chosen in advance. Our choice for k was 10.

- **SVM** (support vector machine)

  SVM translates set of features in a higher dimensional space, where a better separation between features can be achieved. The optimal hyperplane separates space so that distances to the nearest data points are maximized.

- **RF** (random forest)

  Random forest is a state of the art algorithm used for classification and regression problems. It operates on the idea of multiple decision trees construction. The main parameter is the number of trees, which was in our case set to 100.

- **NB** (naive Bayes)

  NB is based on the Bayes' theorem that assumes strong independences between features.

- **NN** (neural networks)

  Deep neural network, inspired by biological nervous systems, is an emerging family of machine learning algorithms which have lately became accessible due to CPU advancements. NN have the ability to be arbitrarily constructed from various kinds of layers. In our context, this method is an exception versus algorithms listed above, because it does not require any feature engineering, as the NN can (with appropriate layers construction) find valuable hidden features by itself on a raw signal.

  More specifically in our case, these layers are 1-dimensional *convolutional* layers and *long short term memory (LSTM)* layer. In this combination, the NN work with sequence related problems in spatial inputs (e.g. signal in time). An input to our implementation of NN is therefore signal phase over some time period (i.e. distance from antennas to users body on e.g. 30 seconds long time window).

  Our NN consists of the following layers (see Figure 3.12):

  1. dropout

     Dropout layers are used to reduce overfitting by randomly choosing neurons and simply ignoring them in later processing. This is a widely used neural network regularization technique to prevent too much adaptation on the training dataset.

  2. 1-D CNN & max pooling

     This layer applies convolutional operation to the input, which greatly reduces memory requirements and leads to more general solutions to vision problems (note that our problem can be presented as 1-dimensional image). Max pooling combines multiple neurons into a single neuron in the upcoming layer.

  3. LSTM layer

     LSTM is a type of recurrent neural network that remembers values over time and it can make predictions based on time series data.

4. fully connected layer & activation

   Fully connected layer connects each neuron on the previous layer to each neuron in continuing layer. Activation neuron outputs final prediction, either to classification or regression problem, depends on the given activation function.

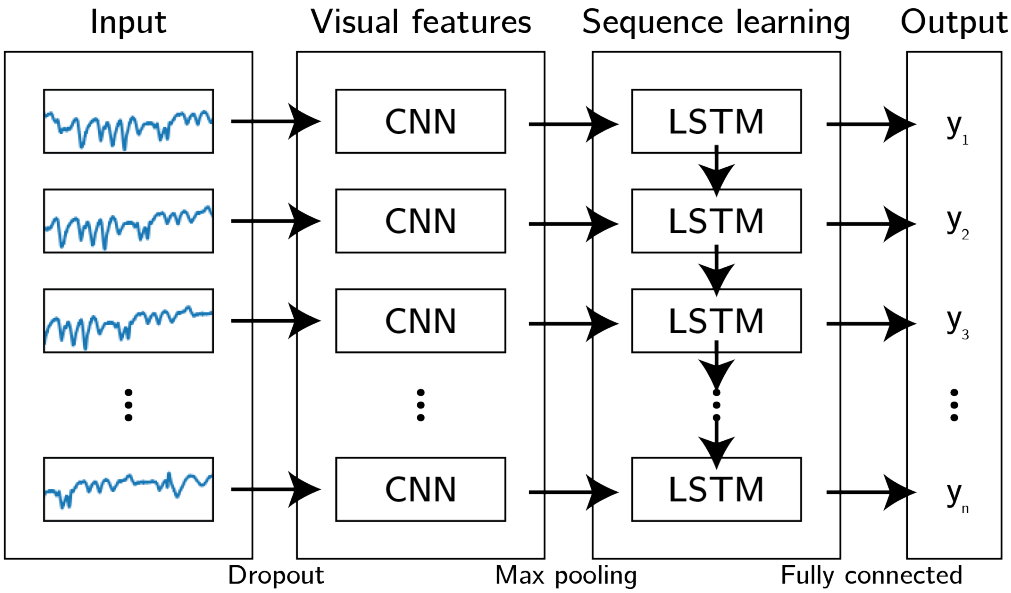

Figure 3.12: Simple example of neural network with the main convolutional and LSTM (long-short term memory) layers. Dropout, max pooling and fully connected layers are presented just to show their positions between the other layers.

Chapter 4

Experimental setup

The goal of Wi-Mind system is to infer cognitive load using wireless signals. To test this system we conducted user study where 23 volunteers were solving cognitive load related tasks in a quiet air-conditioned room while Wi-Mind system was wirelessly acquiring their vital signs. With each participant we collected their demographics, explained the experimental protocol, and had them complete the tasks uninterrupted. Wi-Mind is geared towards sedentary mental task load inference, thus, we collect the data in an office setting with an application Haapalainen et al. constructed to elicit different cognitive load burden [16]. Study protocol diagram for each user can be seen in Figure 4.1.

4.1 Cognitive load application

Cognitive load measurement software was prepared by Eija Haapalainen et al. and tested in cognitive load study in [16]. Martin Frlin adapted this software to Slovenian language. The application runs on a PC and presents the user with six task types:

- Finding hidden pattern (**HP**) – find a given pattern in multiple images;
- Finding A's (**FA**) – choose all words that have a letter "A" in them;

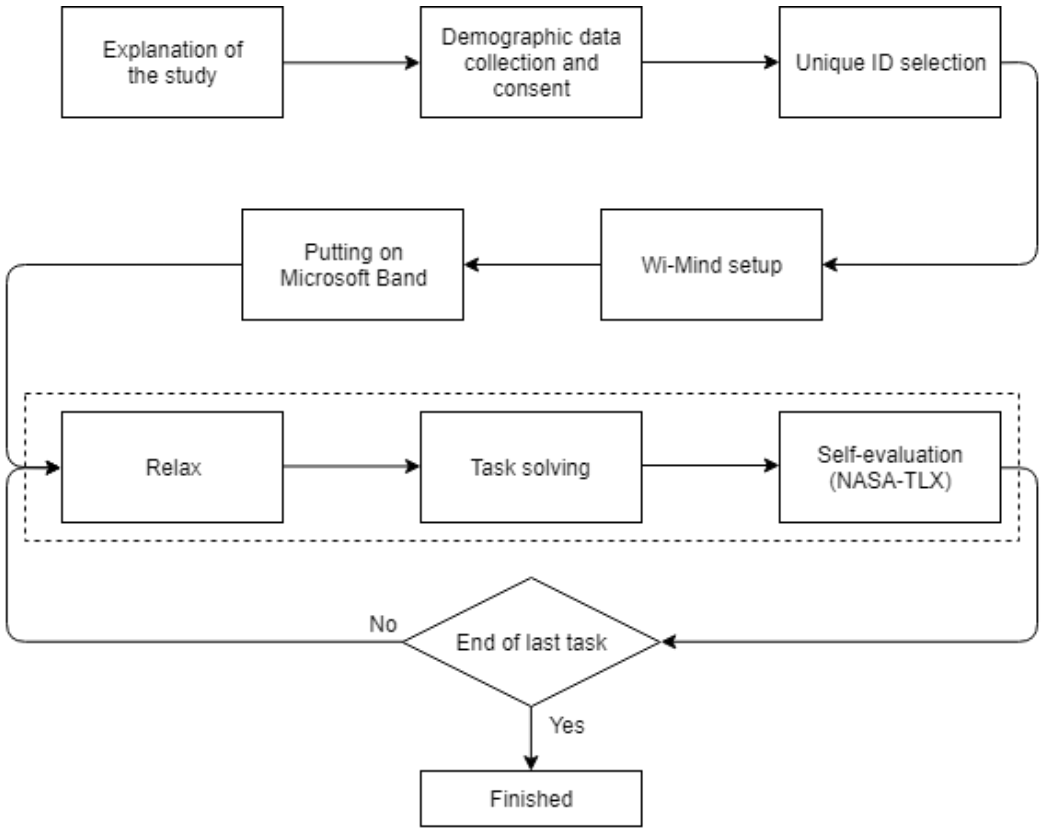

Figure 4.1: User study protocol diagram to test Wi-Mind system.

- Gestalt completion (**GC**) – from a partial image find out what would the whole picture represents and write down the answer;
- Number comparison (**NC**) – in two parallel lists of numbers find those that are equal;
- Scattered X's (**SX**) – in a set of images find letters "X" and click on them;
- Pursuit test (**PT**) – connect values on the left side to the corresponding values on the right side following entangled lines connecting the two sides.

Each of these tasks is presented three times, with three different difficulty levels - *easy*, *medium* and *hard*. While we certainly expect that this objective label already correlates with a person's cognitive load, we also rely on the NASA-TLX questionnaire to infer a person's subjective feeling about the load. The questionnaire, an integral part of the Haapalainen et al.'s app, is presented to users after each of the tasks. All in all there are 18 different tasks (six different task types and three difficulty levels for each one). Each time when a new task is presented, there is also "instructions" slide, for a quick overview of the task. Each difficulty level for one task type has maximum three minutes to solve the task, otherwise the application automatically switches to next slide (i.e. NASA-TLX questionnaire). When one minute is left for solving the task, user is also informed about this with a sound signal. Solving each task requires only a computer mouse, with exceptions of GC and PT, where keyboard has to be used for writing down the answers.

As we can see from the study protocol diagram (see Figure 4.1), each task procedure consists of three intervals - *relax*, *task solving* and already mentioned *self-evaluation*. Just before starting each task, there is a "relax" time frame. This break should be considered as a dividing line between tasks, so that physiological signs between adjacent tasks do not intervene. At the beginning of the first task, this is 1.5 minute long break and in the subsequent

tasks this break is 30 seconds long. After finishing some task or running out of time, NASA-TLX is presented.

NASA-TLX is subjective assessment tool for evaluating one's cognitive load in order to assess a task performance. Although it is a widely used method for assessing mental effort [34], some studies do not consider the self-reports reliable enough to assess cognitive load [56]. NASA-TLX is divided into six subjective question for assessing the:

- mental demand;
- physical demand;
- temporal demand;
- performance;
- effort;
- frustration.

In our case, each of them could be answered with a 5-item scale (e.g. *very low*, *low*, *medium*, *high*, *very high*).

Each time when a user or an application does some action, e.g. click on a button, chooses answers, self-evaluates, runs out of time, timestamp and event is saved to log file. A part of this log file can be seen below:

```
395 ...
396 2018:5:17:17:42:15:874, Number Comparison Question Slide,
      NumberComparison_2.txt, Medium
397 2018:5:17:17:42:21:354, '6312850395-6312850795' selected
398 2018:5:17:17:42:25:619, '1251373807-1251373307' selected
399 2018:5:17:17:42:30:810, '32018591670-32018691670' selected
400 2018:5:17:17:42:39:538, '38210435512-38210535512' selected
401 2018:5:17:17:42:43:682, '35789462806-35789562806' selected
402
403 2018:5:17:17:42:44:411, Rating Slide
404 2018:5:17:17:42:46:610, Mental, Low
405 2018:5:17:17:42:53:346, Physical, Low
```

```
406 2018:5:17:17:43:1:578, Temporal, Low
407 2018:5:17:17:43:4:634, Temporal, Medium
408 2018:5:17:17:43:5:915, Temporal, Low
409 2018:5:17:17:43:7:610, Performance, Very good
410 2018:5:17:17:43:14:450, Effort, Low
411 2018:5:17:17:43:17:338, Frustration, Very low
412
413 2018:5:17:17:43:18:707, Break Slide
414
415 2018:5:17:17:43:48:709, Test Continues Slide
416
417 2018:5:17:17:43:52:19, Instruction Slide, Scattered X's
      Questions
418 ...
```

Log files are later processed again and only relevant features are extracted:

```
...
235 2gu87 NC high 1526571422322 163432 4 0 4 27 True
236 2gu87 NC medium 1526571657466 106945 14 1 14 15 True
237 2gu87 SX low 1526571850374 26429 20 0 20 11 True
...
```

where each row presents data from one task and each column shows (from left to right): task ID, user ID, task label, task complexity, start time (in Unix epoch time), time on task (in milliseconds), number of correct answers, number of incorrect answers, number of all correct answers, TLX (from NASA-TLX questionnaire - this value is calculated by combining/summing all six answers) and boolean value if user finished task (i.e. did not ran out of time). This data is later used in machine learning algorithms, either as meta features or as target variable to predict cognitive load.

4.2 Protocol

In order to make as equal conditions as possible for all participants, we made a brief protocol document. Each individual was stationed in a quiet air-conditioned room. First, we explain what the study is about and briefly

describe Wi-Mind system.

Then, we collect volunteers demographic data (see demographic data form in Appendix A), such as:

- age;
- gender;
- dominant hand;
- level of education.

Finally, each user fills out the consent form. Consent refers mostly to information about the study, that the participation is completely voluntary, that the collected data is used for scientific purposes and the assurance for complete anonymity. Each volunteer is then given a unique ID.

After demographic data collection and ID selection, we setup our Wi-Mind system. As mentioned in Chapter 3, Wi-Mind uses two antennas. As the system is very sensitive, even small movements of the antenna locations can greatly improve our signal accuracy. At this point we start GNU Radio programme and try to improve signal by moving the antennas for a few centimetres back and forth or adjusting the antenna height position.

Another thing, that was not yet mentioned, is Microsoft Band. Microsoft Band (we used Microsoft Band 2) is a smart band for tracking users fitness features, with sensors such as heart rate monitor, gyrometer, galvanic skin response sensor, skin temperature sensor, and more smartwatch-like features [57]. We tell the participants to put on the Microsoft Band on their non-dominant hand (stated in the demographic data) and then we start our custom application on the phone (running Android 4.4.4) to collect the heart rate each second during the study. This is just a reference data to compare this "ground truth" heart rate to heart rate acquired and extracted with Wi-Mind. However, the band is not necessarily 100% accurate, but our decision to use this band was due to its low cost and its software development kit

(SDK), which is available on the internet and allowed us to make custom application for the Android OS.

The last preparation point was made and then we continue with cognitive load study starting our Wi-Mind system and tasks application at the same time. We also tell the users not to make significant body movements or limb motion, just to have clearer wireless signals. While users were solving the tasks, we were in the same room looking at the second screen connected to the same computer they were solving the tasks at. The second screen consisted of the current information about acquired signal (signal phase) and a current slide. If the signal was not clear enough or the user was moving too much, we pointed that to him/her. We also answered all the questions referring to the study tasks.

4.3 Study summary

The study was conducted among 23 volunteers, ageing from 20 to 38, 17 male and 6 female (see Figure 4.2 to see an example of a participant performing the experiment). The average time for completing the experiment was around 45 minutes. An example of the Wi-Mind system signal during the cognitive load study can be seen in Figure 4.3. The crucial step was to find at least 20 volunteers for the study and we did that through personal contacts. All of them participated the study completely voluntary and without any monetary compensation.

Figure 4.2: Cognitive load study for testing Wi-Mind system in action.

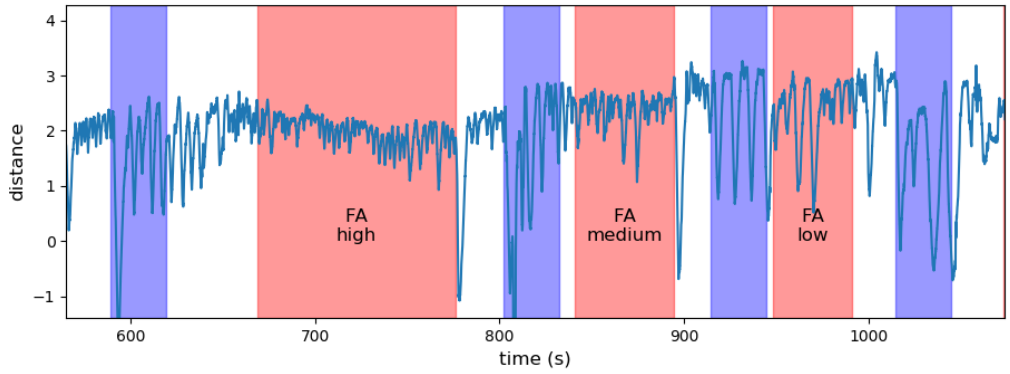

Figure 4.3: Wi-Mind signal through cognitive load study. Red areas present tasks (e.g. FA – finding A's) with different difficulty levels, blue areas present relax time frames and white areas present either solving the NASA-TLX questionnaire or preparation for the task.

Chapter 5

Results

While we collected all the relevant data from our study users, here we present our findings about users cognitive load engagement using wireless signals. In this chapter we showcase the micro benchmark for heart rate accuracy, just to see how accurate Wi-Mind system is in extracting heart rate during the study, we present some of descriptive statistics regarding our study and show prediction results for different cognitive load engagements.

5.1 Heart rate benchmark

In Chapter 4 we mentioned that we used Microsoft Band to measure "ground truth" heart rate during the study. We used this data to compare Wi-Mind's extracted heart rate. Both of the time series (Wi-Mind heart rate and Microsoft Band heart rate, i.e. series of heart rates through the whole study) were aligned in time and compared using mean absolute error (MAE) and mean squared error (MSE) metrics (see Equations 5.1 and 5.2) for each user. Some of the best and worst per-user metric values can be seen in Table 5.1.

$$MAE = \frac{\sum_{i=1}^{n} |Y_i - \hat{Y}_i|}{n} \qquad (5.1) \qquad MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \qquad (5.2)$$

Table 5.1: Error metrics comparing extracted Wi-Mind and Microsoft Band heart rate for some of the study users.

user ID	MAE	MSE
pmyfl	4.935	52.988
td5pr	5.218	57.331
fzchw	5.892	57.477
c24ur	6.155	73.143
⋮	⋮	⋮
62i9y	15.852	385.736
ctsax	15.929	408.712
1mpau	17.038	428.921
r89k1	20.935	555.897

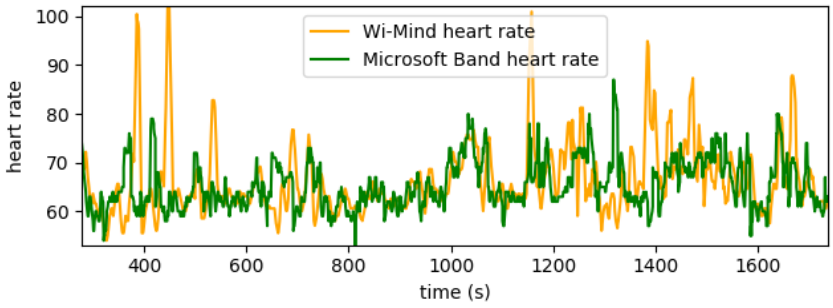

Figure 5.1: Comparison of Wi-Minds extracted heart rate and Microsoft Band heart rate for one of the study users.

An example of heart rate through time can be seen in Figure 5.1. In this example we see some outlier heart rate values. The reason for this is usually moving hands from and to the computer keyboard, which makes wireless signal reflections almost indistinguishable, which produces wrong heart rate extraction. Another comparison of the heart rate can be seen in Figure 5.2, where each box plot shows heart rate for one user acquired or extracted with band and Wi-Mind, respectively. Wi-Mind extracted heart rates are far from perfect, however, trend on the right plot looks like it is sorted ascending by

median values, similar as to left plot (box plots are sorted by median values from the left plot and each user box plot on the right corresponds to the same position box plot on the left). Note that third user from the right on the band heart rate plot looks like it was static all the time. This is a result of error in capturing the data from Microsoft Band, as there were missing values and this does not correspond to the actual heart rate during the whole study.

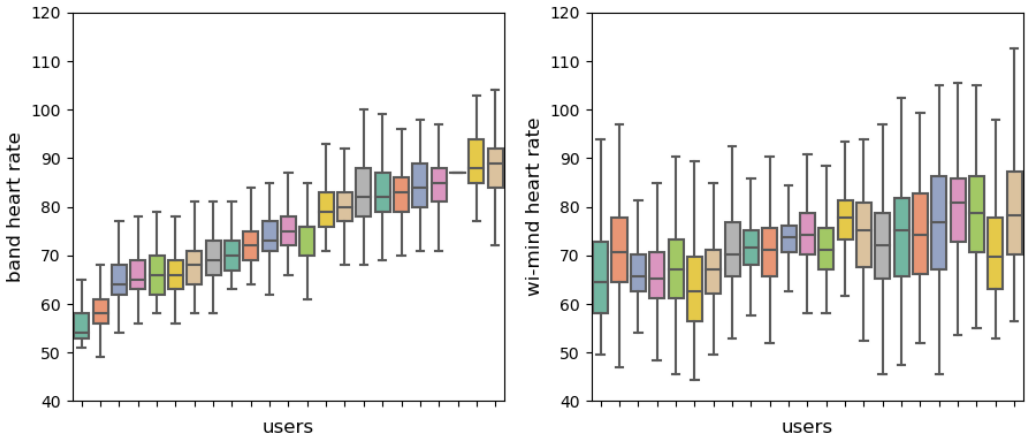

Figure 5.2: Box plot comparison of Microsoft Band acquired heart rate and Wi-Mind extracted heart rate during the whole study for all users. Each user on the right corresponds to the same user on the left.

5.2 Descriptive statistics

In order to better understand the data about the tasks and cognitive load application in general, we examined descriptive statistics about the participants' performance. This information might help us to deal with outliers and similar data separations in the learning/prediction phase.

As mentioned in Chapter 4, users in the study were exposed to three types of task difficulties (*low*, *medium* and *high*). The most intuitive hypothesis on the relationship between the task difficulty and the time a user spent on

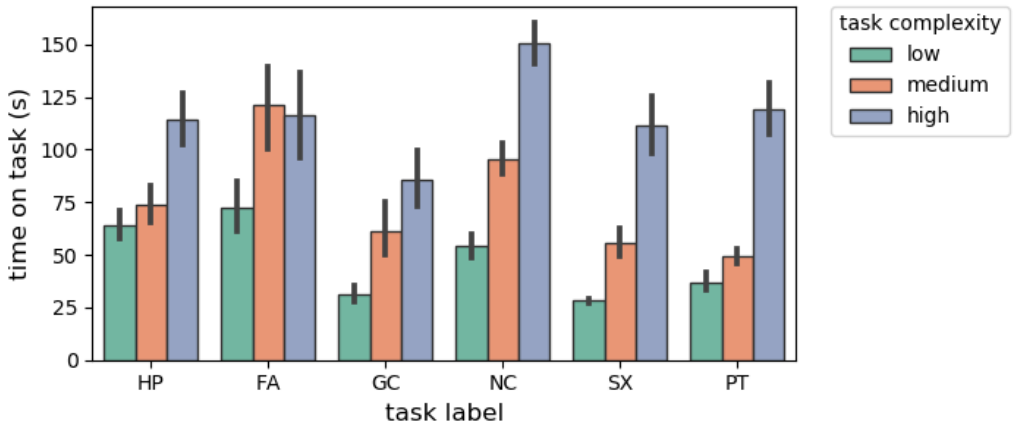

Figure 5.3: Bar plot presenting time to finish the task grouped by task label and task complexity. Black lines present 95% confidence interval.

the task is that tasks of higher difficulties would need more time to finish the task. This holds true in this study (see Figure 5.3), however, there is clearly some overlapping between tasks of low and medium difficulty and minor overlapping of medium and high difficulty tasks. Interestingly, on average users took task FA of medium difficulty longer to finish than the highest difficulty. Same findings can be seen for average TLX value in Figure 5.4. Again, median TLX values do correlate with task difficulties, yet overlapping in TLX between adjacent difficulties stays or even increase. Both overlapping between task difficulties (in time and TLX) might indicate not so distinct separations between the difficulties and that users did not really see the difference between e.g. low and medium tasks. Another possibility would be the task type being easy by itself (e.g. HP task is easier than NC, in theory). HP task type has been found to be less difficult in overall than the other task types. This tells us that using the same machine learning model for all tasks is unlikely to be able to discern between tasks of different nominal complexities.

To evaluate task difficulties, instead of looking at subjective TLX measure, we can also look at the more objective measure – the number of incorrect

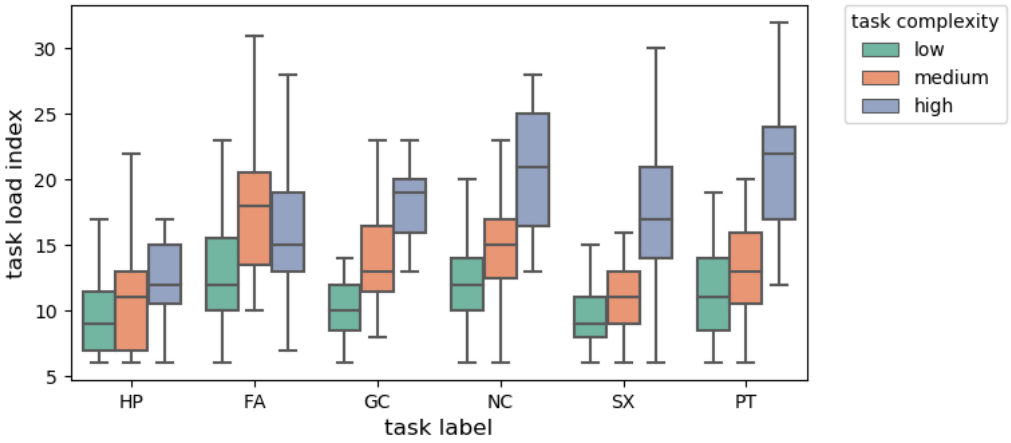

Figure 5.4: Box plot presenting task load index grouped by task label and task complexity (outliers are not shown).

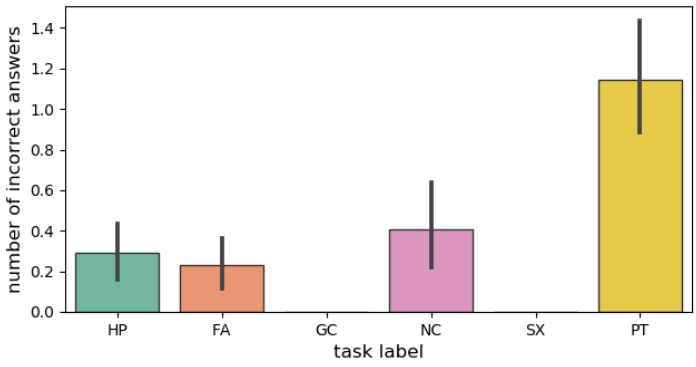

Figure 5.5: Box plot presenting an average number of incorrect answers per task type.

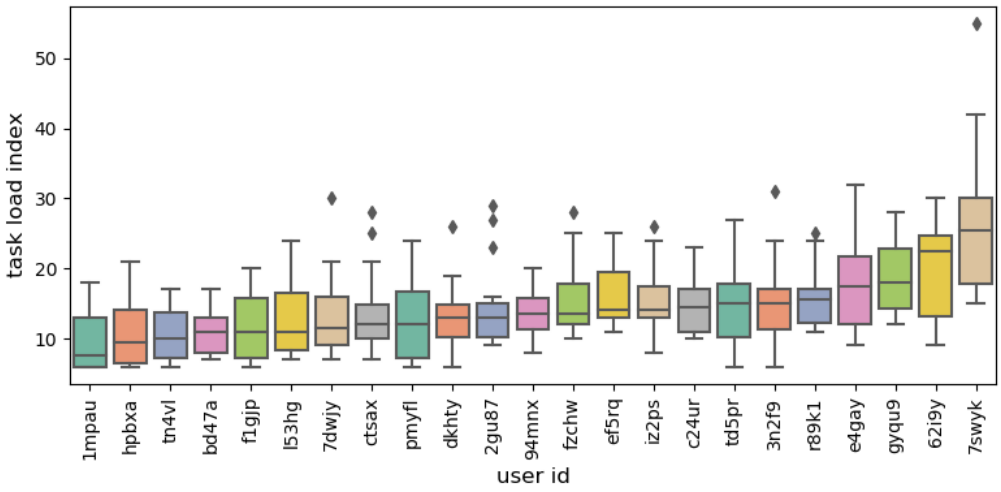

Figure 5.6: Box plot presenting task load index by user id.

answers per task type. This can be seen in Figure 5.5, where we can count out the GC and SX task types, because we did not count incorrect answers, as we did not have ground truth values. The easiest task types (looking at this objective measure) are therefore FA and HP. This holds true for HP, as HP was considered as the easiest (in overall) also with the subjective TLX measure. Task type with the most incorrect answers overall is PT. This is somehow expected, as answers in one task slide are interdependent with each other, i.e. if you make one mistake, this can propagate to other mistakes.

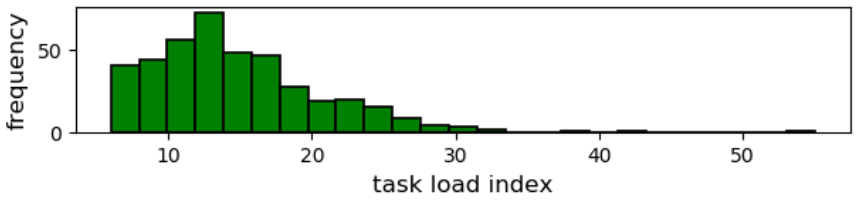

Figure 5.7: Histogram showing overall task load index distribution.

Some of the users might find all tasks easy to solve (or difficult). This can be seen in Figure 5.6, where box plot shows that some users were subjectively

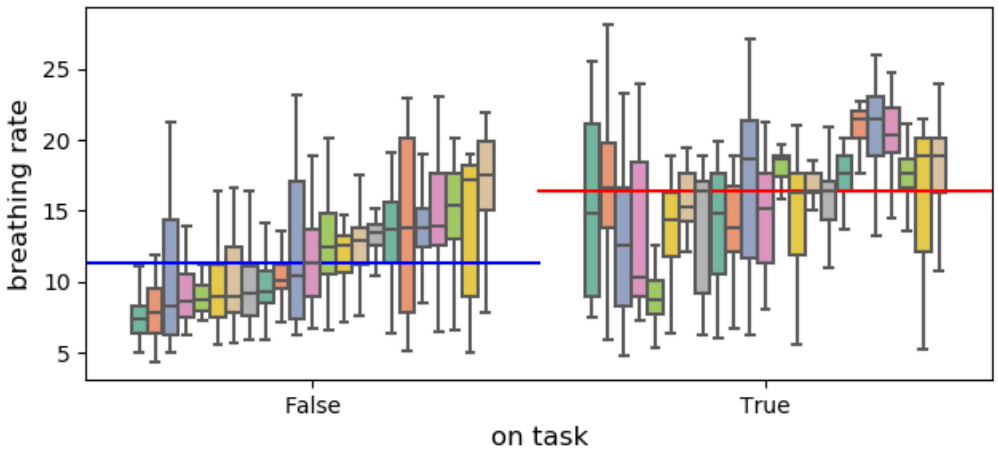

Figure 5.8: Box plot presenting each user breathing rates, extracted with Wi-Mind, grouped by *on task* meta-feature – representing whether a user is relaxing or solving a task (outliers are not shown). Each user on the right corresponds to the same position on the left side. Blue and red line show overall median value for a specific class.

evaluating at different span scale. E.g. user with ID *7swyk* (far right) has, in comparison to other users, far higher TLX values and span. This could be a result of not understanding the questionnaire, having different personality or even having different level of education, but in our case the latter can be excluded since there were no extreme education level differences between the users. To deal with this kind of "extreme" values, we can just cut out these kind of users when predicting TLX, as a joint model is unlikely to work well for all users. In Figure 5.7 we see the overall distribution of TLX values and clearly most of the TLX values fall in the 5-20 range.

In Figure 5.8 we presented breathing rates, extracted with Wi-Mind while relaxing and solving a task. Again, each user box plot on the right corresponds to the same position box plot on the left and is sorted by ascending median values from the breathing rates when relaxing (left). Overall, it does look like breathing rates increase when solving a task, however, there are still some exceptions, where extracted breathing rate does not change signifi-

cantly when user is relaxing versus solving a task. Similar findings of multiple studies were reviewed in [5], where they detected increase in breathing rate while solving mentally demanding tasks.

5.3 System evaluation

To evaluate Wi-Mind system we used Orange [17] and Keras [18]. Orange is a popular data mining toolkit, workflow based software, with predefined widgets for preprocessing, evaluating, classification, regression, visualization etc. Keras is a Python deep learning library for neural networks construction, which is running on top of TensorFlow, an open source software for numerical computation.

We are interested to build a general purpose cognitive load inference system, i.e. no matter who is using the system. Because of this assumption, the most appropriate way to evaluate Wi-Mind system is with the *leave one person out* validation. This means that the learning phase is done on all users data, except one. Testing/evaluating (accuracy calculation) is done on this one user data. The whole procedure is then repeated for every person and the final accuracy is averaged from the before-mentioned one user accuracies. However, since we are dealing with diverse data (e.g. different TLX values – see Figure 5.6, breathing rates – see Figure 5.8, etc.) and not so many user instances, instead of leave one person out validation (on all users), we also try to validate Wi-Mind system on a group of similar users.

Most of our results were tested with k-NN, SVM, RF, NB and majority classifier/mean value as the baseline. Where mentioned, we also used the deep NN (1-D CNN + LSTM). The latter was, instead of being done in Orange as all the other approaches, evaluated in Keras. NN is being used without the extracted features, instead it works directly on the raw data. As mentioned in Chapter 3, we also introduced the meta-feature to see whether some time window is too noisy or not. Where mentioned we used this meta-feature to remove noisy data.

5.3.1 Relaxation/task solving classification

In the preliminary step we were curious to see whether the acquired data can be used to discern between a person being busy and resting. To evaluate such a basic classifier, we divide the data into *relaxing* (intervals when a user is instructed to relax) and *busy* (last 30 seconds while solving task) time frames (see Figure 5.9 to get the idea for relax/busy intervals).

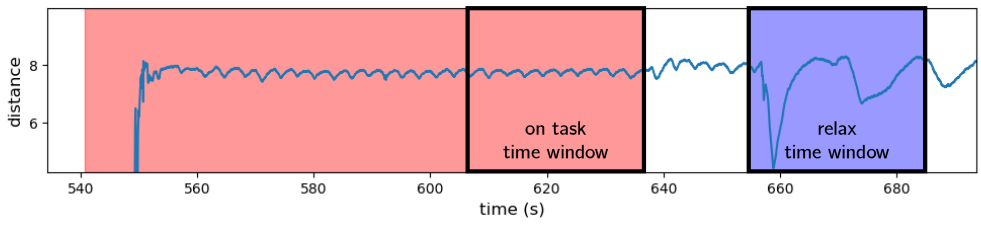

Figure 5.9: Raw signal with *on task* (red) and *relax* (blue) time frames. Black rectangles present time windows that were used in classification between the two.

The relax/busy time frames classification results can be seen in Table 5.2. Algorithms with the highest classification accuracies in non-filtered dataset are RF and NB. Similar results can be seen in a noise filtered data, where both of the algorithms still have the highest accuracies, just above 70%. We see noise filtering meta-feature does not significantly improve accuracies. Without directly extracting features (using deep neural network and raw signal), we get even better results with a CA of 75.2%.

Table 5.2: Leave out person out validation for relaxation/task solving classification. First half of the table uses all data time frames (without noise filtering) with class instances ratio *relaxed*:*on task* 414:414 and second half (with noise filtered data) ratio of 313:338. Last row shows NN accuracy.

Method	AUC	CA
k-NN (k=10)	0.703	0.668
SVM	0.651	0.571
Random forest	0.794	**0.705**
Naive Bayes	0.773	0.699
Majority	0.5	0.5
k-NN (k=10)	0.696	0.650
SVM	0.642	0.590
Random forest	0.788	0.704
Naive Bayes	0.779	**0.707**
Majority	0.5	0.519
1-D CNN + LSTM		**0.752**

Top three features, that contributed the most to these results, are breathing rate, standard deviation of inter-breath intervals, and mean value of inter-breath intervals (according to information gain scoring method) – see Table 5.3. The first two are correlated, since they represent similar feature calculated in two different domains (frequency domain and time domain).

Table 5.3: Top five information gain scores for busy versus relax classification problem.

Feature	Information gain
br_rate	0.121
br_IBI_mean	0.116
br_IBI_std	0.097
br_freq_6_12	0.075
br_raw_rms	0.051

Personalized tests

If we take into consideration personalized tests (i.e. predicting classification with only one user data) and validate them using leave one out validation, we get the following results (see Table 5.4):

Table 5.4: Classification accuracies into relax/busy classes with personalized modelling. Class instances ratio (busy:relax) for each user is 18:18 and leave one out validation is used. Last column shows average values over all user accuracies for specific method.

Method	*1mpau*	*2gu87*	*3n2f9*	...	*r89k1*	*td5pr*	*tn4vl*	average
k-NN (k=10)	0.5	0.778	0.306	...	0.583	0.694	0.472	0.604
SVM	0.639	0.833	0.667	...	0.639	0.806	0.833	0.721
RF	0.694	0.861	0.611	...	0.667	0.833	0.694	0.721
NB	0.694	0.833	0.583	...	0.75	0.833	0.722	0.734
Majority	0.5	0.5	0.5	...	0.5	0.5	0.5	0.5

Overall it does look improved, but for each model we are dealing with only 36 different time windows/instances, which can lead to small dataset problem. More generalized results could be achieved with more instances for each user.

Feature normalization

As we noticed, different users have different breathing rates at e.g. relax time frames. Although there are differences at individual user when relax vs. solving a task times (each user should have higher breathing rate when solving a task), this cannot hold when comparing breathing rates from different users, as people have different fitness levels and current states (e.g. one users' breathing rate at *task solving state* can be similar as breathing rate of another user at *relaxing state*).

This brings us to "feature normalization", where we try to modify the *breathing rate* feature in our dataset. The modification puts every users

breathing rate at relax time frames to the approximately same rate. Since we have only 23 users, we try to get better results with modifying such features, without changing the ratios between each feature values.

Table 5.5: Leave out person out validation for relaxation/task solving classification with normalized *breathing rate* feature. Class instances ratio *relaxed*:*on task* is 414:414.

Method	AUC	CA
k-NN (k=10)	0.752	0.704
SVM	0.67	0.58
Random forest	0.806	**0.746**
Naive Bayes	0.78	0.723
Majority	0.5	0.5

From Table 5.5 we see that normalizing *breathing rate* feature does improve classification accuracy, since in this case users have similar breathing rate while relaxing.

5.3.2 Change in task engagement classification

Next, we try to predict the cognitive load increase/decrease. User should go into the phase of decreasing task engagement while transitioning from the task solving state into the relaxing state and vice versa (see Figure 5.10). As this change does not happen suddenly, but gradually, we took multiple time windows (more specifically 10 sliding time windows for each class). With this method we also try to make our dataset more robust, as small datasets might lead to wrong conclusions.

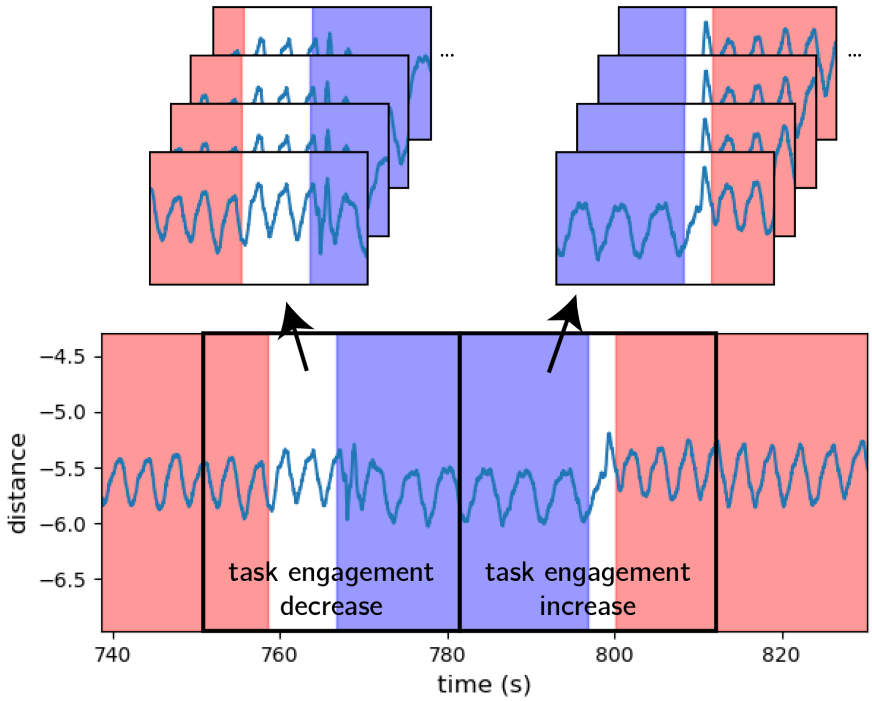

Figure 5.10: Signal presentation of the sliding window idea for the change in task engagement classification. First half presents decrease and second one increase in task engagement. Each sliding window features are marked as one instance and have suitable final class (i.e. *decrease* or *increase*).

Table 5.6: Leave out person out validation for increase or decrease of task engagement classification. First half of the table uses all data time frames (without noise filtering) with class instances ratio *decrease*:*increase* of 4301:4301 and second half (with noise filtered data) ratio of 3635:3719.

Method	AUC	CA
k-NN (k=10)	0.671	0.622
SVM	0.517	0.508
Random forest	0.724	**0.664**
Naive Bayes	0.716	0.643
Majority	0.5	0.5
k-NN (k=10)	0.668	0.622
SVM	0.538	0.530
Random forest	0.726	**0.680**
Naive Bayes	0.705	0.644
Majority	0.5	0.506
1-D CNN + LSTM		**0.681**

From Table 5.6 we see that RF slightly overtakes with CA of 66.4% and 68% in non-filtered data and filtered data, respectively. Bottom row in this table shows classification accuracy using deep NN. It is comparable with random forest classification with a value of 68.1%.

Top two features, that contributed to these results the most, are the breathing rate difference (i.e. breathing frequency from the second half of a time window substituted by breathing frequency from the first half) and mean value of the inter-breath intervals (according to information gain score) – see Table 5.7.

Table 5.7: Top five information gain scores for task engagement increase versus decrease classification problem.

Feature	Information gain
br_IBI_mean	0.062
br_rate_difference	0.057
br_rate	0.029
br_freq_18_24	0.017
br_raw_std	0.010

Similar users test

Our assumption for task engagement increase/decrease was that people might start breathing faster/slower when users are introduced to a new task or are instructed to relax. But since users have different breathing rate changes (as seen in Figure 5.8), we divided them into two groups, where one includes users with small changes in extracted breathing rates (i.e. average breathing rate difference between relax and on task time frames is less than five – *group 1*) and bigger changes in extracted breathing rates (average breathing rate difference between relax and on task time frames is more than five – *group 2*) – this division makes two groups of approximately same number of users:

- *group 1*: *pmyfl*, *ef5rq*, *62i9y*, *dkhty*, *1mpau*, *e4gay*, *c24ur*, *td5pr*, *r89k1*, *l53hg*, *f1gjp*, *94mnx*

- *group 2*: *hpbxa*, *7dwjy*, *bd47a*, *7swyk*, *tn4vl*, *ctsax*, *2gu87*, *fzchw*, *gyqu9*, *3n2f9*, *iz2ps*

The classification accuracies for each group can be seen in Table 5.8. We see that *group 2* has improved accuracies overall, since algorithms detect bigger changes in breathing rates more distinctly.

Table 5.8: Classification accuracies into engagement increase/decrease classes with grouped users modelling. Class instances ratio for *group 1* is 2376:2376 and *group 2* 2178:2178. Leave one person out validation was used for each group.

Method	*group 1*	*group 2*	*average*
k-NN (k=10)	0.558	0.657	0.608
SVM	0.486	0.530	0.508
RF	0.563	**0.695**	0.629
NB	**0.579**	0.686	0.633
Majority	0.5	0.5	0.5

5.3.3 Task complexity classification

Another thing to consider is task complexity classification. We do not expect good results, as a well trained user might find all of the given tasks easy thus preventing us from distinguishing the nominal task difficulty merely on the physiological signals. To evaluate such a multi-class classifier, we constructed a dataset taking 10 instances of time frames from each task difficulty and then removed too noisy instances with our meta feature. Instances ratio can be seen in Table 5.9, as well as classification accuracies. RF again shows the highest CA, but the accuracy is still close to baseline. If we take out the *medium* difficulty from the dataset, we get no big improvement.

Slightly higher CA can be seen at RF classifier. The reason for this could probably be small body movements/limb motion to reach out the mouse or keyboard at some task types. As mentioned, Wi-Mind is very sensitive and even the smallest irregular movement could be seen as higher amplitude changes in the acquired signal. Because of this assumption we made per-task models (see Table 5.10 and Table 5.11 for three and two class problem, respectively).

Table 5.9: Leave out person out validation classification for overall task complexity (class instances ratio low:medium:high is 1111:1159:1134) on the left and classification for low and high task complexities only on the right (instances ratio low:high is 1111:1134).

Method	AUC	CA
k-NN (k=10)	0.513	0.343
SVM	0.521	0.328
Random forest	0.55	**0.369**
Naive Bayes	0.502	0.337
Majority	0.5	0.34
1-D CNN + LSTM		0.334

Method	AUC	CA
k-NN (k=10)	0.531	0.534
SVM	0.460	0.528
Random forest	0.533	**0.520**
Naive Bayes	0.517	0.5
Majority	0.5	0.503
1-D CNN + LSTM		0.501

The results show that per-task models give better results. The highest classification accuracies can be seen at GC and NC task type. One assumption for higher GC accuracy is that users may write on keyboard at different speeds, while solving different task complexities (of the same task), which results in different noise in signal. However, NC accuracy is not so straightforward. The feature, that influenced the most at the NC model, is *HRV high frequency* (according to information gain score – see Table 5.12). HRV high frequency value is usually decreased under time pressure [58] (note that users took NC task at the highest difficulty the longest time to solve – see Figure 5.3 earlier in this chapter). Another important thing to mention is that HRV frequency features are calculated over a 100 second long time window, since shorter times for HRV frequency features are almost impossible to calculate even with high precision equipment (note that we still use the same number of time window instances as e.g. extracted heart rate – the difference is only the extracted time window length). Since NC is the task that had the longest *time on task*, these HRV frequency features could influence on the results only on tasks with the highest *time on task* value.

Table 5.10: Classification accuracies of task complexities by task (i.e. per task models – each model was made on one task type). Class instances ratio for each model is 230:230:230 (low:medium:high). Leave one person out validation was used for each task type.

Method	FA	GC	HP	NC	PT	SX
k-NN (k=10)	**0.365**	0.368	0.29	**0.442**	0.301	0.339
SVM	0.322	0.294	**0.356**	0.281	**0.374**	0.301
Random forest	0.301	0.397	0.343	0.433	0.310	**0.381**
Naive Bayes	0.287	**0.401**	0.346	0.41	0.345	0.337
Majority	0.333	0.333	0.333	0.333	0.333	0.333
1-D CNN + LSTM	0.358	0.327	0.331	0.295	0.335	0.299

Table 5.11: Classification accuracies of task complexities by task (i.e. per task models – each model was made on one task type) with only low and high task complexities. Class instances ratio for each model is 230:230 (low:high). Leave one person out validation was used for each task type.

Method	FA	GC	HP	NC	PT	SX
k-NN (k=10)	**0.533**	0.528	0.451	0.581	0.48	**0.591**
SVM	0.464	0.466	0.565	0.507	0.568	0.473
Random forest	0.429	**0.601**	**0.575**	**0.654**	**0.577**	0.556
Naive Bayes	0.441	0.57	0.541	0.534	0.504	0.501
Majority	0.5	0.5	0.5	0.5	0.5	0.5
1-D CNN + LSTM	0.541	0.492	0.532	0.52	0.495	0.51

Table 5.12: Top five information gain scores for task complexity (*low* – *high*) classification problem for NC task type.

Feature	Information gain
hr_HRV_hf	0.129
hr__HRV_lf_hf	0.068
br_freq_24_30	0.062
hr_RMSSD	0.061
br_rate_difference	0.060

5.3.4 Task load index regression

Instead of trying to predict task difficulty, we also try to predict subjective label – TLX (see Table 5.13 for all and only low – high task difficulties error metrics). MSE and MAE are larger that baseline errors, which indicates that our models are not capable of predicting the TLX.

Table 5.13: Leave one person out validation regression on all user data for overall task load index on the left (number of instances is 4140) and for low and high task complexities only on the right (number of instances is 2760).

Method	MSE	MAE
k-NN (k=10)	52.556	5.527
SVM	60.718	6.171
Random forest	47.868	5.313
Mean	41.431	4.896

Method	MSE	MAE
k-NN (k=10)	59.295	5.941
SVM	58.951	5.921
Random forest	51.475	5.659
Mean	45.887	5.215

Table 5.14: Leave one person out validation for task load index regression on a group of similar users (group 1) without medium difficulty tasks. Number of instances is 720.

Method	MSE	MAE
k-NN (k=10)	29.278	4.315
SVM	30.245	4.424
Random forest	29.013	4.395
Mean	25.119	3.958
1-D CNN + LSTM	28.254	4.241

Table 5.15: Leave one person out validation for task load index regression on a group of similar users (group 2) without medium difficulty tasks. Number of instances is 960.

Method	MSE	MAE
k-NN (k=10)	34.445	4.692
SVM	33.344	4.602
Random forest	37.106	4.796
Mean	25.863	4.085
1-D CNN + LSTM	24.881	4.149

To take into account similar users (i.e. users who evaluated some tasks with the similar self-evaluation scores), we ran tests on two groups (see Tables 5.14 and 5.15). Groups consist of the following user IDs:

- group 1: *7dwjy, bd47a, f1gjp, hpbxa, l53hg, tn4vl*;
- group 2: *94mnx, fzchw, ef5rq, iz2ps, c24ur, td5pr, 3n2f9, r89k1.*

To divide users in these groups, we helped ourselves with the box plot shown earlier in this chapter (see Figure 5.6). Division does not include extreme values. The results still do not show big improvement in error metrics. The best method is still NN, which is the closest to the error of mean method, but it does still have higher error that mean value.

To simplify our problem, we made models on only GC, NC and PT task types (each one separately) of difficulty levels low and high, as there is clearly easier distinction between TLX values between the two (see Figure 5.4). The results for each group can be seen in Table 5.16 and Table 5.17. Far from perfect and a with a smaller dataset, we get expected results. One exception, which can be seen in both tables, is at task type PT, where RF performs slightly better than baseline regressor. The most decisive features at PT task type are breathing related features – breathing rate and median value from the raw signal.

Table 5.16: Leave one person out validation for task load index regression (group 1) on specific tasks (left GC, center NC and right PT) with only low and high task difficulties. Number of instances is 120.

Method	MSE	MAE
k-NN	30.33	4.727
SVM	38.672	5.079
RF	33.942	5.106
Mean	27.147	4.5

Method	MSE	MAE
k-NN	19.161	3.742
SVM	**14.675**	**2.974**
RF	21.504	3.818
Mean	16.142	3.467

Method	MSE	MAE
k-NN	61.617	6.718
SVM	41.16	5.261
RF	**29.715**	**4.449**
Mean	37.86	5.383

Table 5.17: Leave one person out validation for task load index regression (group 2) on specific tasks (left GC, center NC and right PT) with only low and high task difficulties. Number of instances is 160.

Method	MSE	MAE
k-NN	33.06	4.551
SVM	31.684	4.858
RF	28.094	4.417
Mean	21.12	4.0

Method	MSE	MAE
k-NN	42.634	5.784
SVM	31.295	4.855
RF	42.337	5.960
Mean	25.793	4.455

Method	MSE	MAE
k-NN	31.5608	4.577
SVM	30.689	4.812
RF	**20.373**	**3.788**
Mean	24.909	4.509

5.3.5 Neural network approach

Since neural network method requires some parameter tweaking, we introduced this section, where we present our approach towards some of the classi-

fication problems mentioned in previous sections using neural network. The difference between this approach and the others (i.e. k-NN, SVM, RF, NB), is that this works on a raw signal, whereas the rest of them work on extracted features.

Parameters, that can be modified in our neural network are (see Figure 5.11)

- convolutional layer parameters
  - kernel size
  - number of filters
  - stride
- pool size in max pooling layer
- number of units in LSTM layer
- dropout layer rate (set to 0.25)

In the following tables we were changing most of the above mentioned parameters in different scenarios, just to see how well does NN perform in the given problems. We did not change dropout rate, as we think this is a core layer to overcome overfitting and was set to an optimal value of 0.25.

First classification problem is relaxation/task solving problem (i.e. determine if user is solving a task or relaxing). Our static parameters values are *number of filter = 64*, *stride = 2*, *pool size = 4* and *number of LSTM units = 256*. Changing the *kernel size* parameter does improve CA in this problem. In the Table 5.18 we see that lowering this parameter can bring us to CA of 75.2% in 100 number of epochs. Smaller *kernel size* also decreases time to build such a network, as it reduces computational space with shorter "time windows".

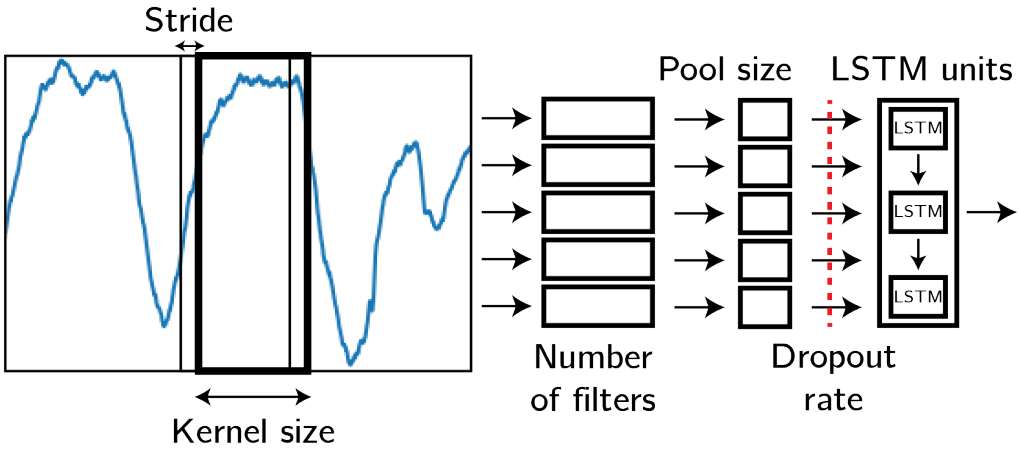

Figure 5.11: Overview of modifiable parameters in our neural network.

Table 5.18: Leave one person out validation with neural networks for relaxation/task solving classification through epochs. Class instances ratio *relaxed*:*on task* is 358:358 (each instance presents 30 second long time window). Convolution layer parameter – kernel size is set to 200 (left), 20 (middle table) and 10 (right).

# epoch	average CA
10	0.612
20	0.653
30	0.674
40	0.674
50	0.672

# epoch	average CA
20	0.676
40	0.702
60	0.718
80	0.727
100	0.726

# epoch	average CA
20	0.676
40	0.715
60	0.724
80	0.748
100	0.752

Classifying into task engagement increase and decrease problem with static values of *kernel size* = *10*, *number of filters* = *64*, *stride* = *4*, and *number of LSTM units* = *256* brings us to the Table 5.19. Increasing *pool size* does not improve CA. However, changing the number of *filters* to 32 and *kernel size* to 100 does improve our CA (see Table 5.20) and is comparable with the CA calculated with previous methods (see Table 5.2).

Table 5.19: Leave one person out validation with neural networks for task engagement increase/decrease classification through epochs. Class instances ratio increase:decrease is 3910:3910 (each instance presents 30 second long time window). Pool layer parameter is set to 4 (left) and 8 (right).

# epoch	average CA
10	0.682
20	0.67
30	0.665
40	0.65
50	0.651

# epoch	average CA
10	0.602
20	0.597
30	0.600
40	0.606
50	0.600

Table 5.20: Leave one person out validation with neural networks for task engagement increase/decrease classification through epochs. Class instances ratio increase:decrease is 3910:3910 (each instance presents 30 second long time window). Pool layer parameter is set to 4, kernel size to 100 and number of nodes in convolutional layer set to 32.

# epoch	average CA
10	0.662
20	0.671
30	0.681

Chapter 6

Discussion

The goal of this work was to infer human cognitive load using wireless signal. We conducted a study with 23 volunteers solving tasks of different complexities. Our first idea was to extract heart beats from wireless signal as it has been shown that heart beats or heart rate variability correlate with an individual's cognitive load. We attempted to extract heart beats from the wireless signal, but were not as accurate as the Vital-Radio authors [9]. Our error metrics (comparing Wi-Minds extracted heart rate with Microsoft Band heart rate) were showing relatively high values. One difference between ours and Vital-Radio article's evaluation section article is that they used a chest band as a reference for ground truth heart rate, where our reference point for heart rate was acquired with Microsoft Band. In the article the authors mentioned that the accuracy exceeded 98% for heart rate, even when users were using a phone or a laptop with daily activities. However, in their experiment users were engaged in daily activities, while extracting heart rates, for 5 minutes, where our user study lasted for cca. 45 minutes. Besides extracting low accuracy heartbeats we also extracted breathing of a person. We did not have ground truth respiratory measure, but were more self-confident into acquiring and extracting these values, as breathing is significantly more apparent in a visualised raw signal than heart beats are. Another comparison of our system with Vital-Radio brings us to the cost of the system –

Wi-Mind system uses USRP B210 connected with USB port directly to the computer, whereas Vital-Radio uses additional hardware, allowing it to make more precise vital sign monitoring and higher range resolution. We managed to make similar vital sign monitor with cheaper hardware.

Evaluating the results of our system shows limitations of using this kind of a radar for cognitive load inference. Previous work was relying into different kinds of equipment, such as smart watches [37], ECG armband, eye tracker and EEG head set [16]. It has been shown that ECG measures can indeed correlate with differences in cognitive load. We did not achieve as accurate of results, but we still got performance values above baseline for some problem definitions. This was expected as we did not use highly accurate intrusive equipment. One of the possible flaws of radar-based vital sign monitor is that while it can measure vital signs when a person is not making excessive limb motion, it has difficulties to do so when a user is moving a computer mouse or produces even higher wireless signal noise when using a keyboard.

While we did manage to predict relax vs. busy states and transitions between them, to some extent, we also tried to infer cognitive load while comparing different tasks. Features, that contributed to most to the positive results of our study, were mostly related to breathing (i.e. average breathing rate, standard deviation of inter-breath intervals, mean value of inter-breath intervals, difference between breathing rates from first and second half in one time window). Breathing, however, is rarely used in cognitive load inference research as it is difficult to measure with commodity wearable devices, such as fitness wristbands. Using Wi-Mind in symbiosis with wearables is an exciting avenue for future research. Per task modelling brought promising results, where *number comparison* task type had the accuracy of 65% with the *HRV high frequency* being the most influencing feature to distinct between two task difficulties. Users required significantly more time to solve this type of a task, which could join with the findings that *HRV high frequency* feature is usually decreased under time pressure [58]. Immediate improvements could include testing Wi-Mind with a higher number of volunteers or with users

with different age/physical fitness, in order to make our dataset bigger and more representative. Features we used are based on experienced intrusive means of measuring vital signs. Collected wireless signal phase data might hide additional features potentially related to cognitive engagement. Being feature oblivious, we also tried a convolutional neural network approach which led to slightly better results in some cases, but was not showing major improvement. To sum up – some of the possible future work improvements could include:

- more study subjects;
- diverse study subjects (users from different generations and/or different physical fitness);
- easier separation between the task complexity levels;
- longer tasks, so that HRV frequency features can be extracted more confidently;
- use Markov chains to build a model for task engagement to see whether the changes between different states happen.

Chapter 7

Conclusion

Modern technology, although facilitating our daily tasks, can have negative consequences on our work by making inconsiderate interruptions, which can reduce our work performance or induce stress. As we live in the world of pervasive technology enabling always-on connectivity, improving human-computer interaction is a critical issue. This interaction could be improved with the machine knowing ones cognitive load by lowering the number of unnecessary interruptions. Cognitive load correlates with humans physiological signs, yet, inferring it requires advanced approaches in human monitoring, signal processing, and machine learning.

In this thesis we investigate a wireless radar based technology to infer cognitive load by observing users vital signs. We used an implementation of a radar to build a Wi-Mind system that measures distance to users body, which translates to vital signs, such as breathing and heartbeat signals. We engineered features according to best practices from the literature and extracted them from the collected wireless data. To test our system, we conducted a study on 23 volunteers while solving tasks of different difficulties. With the acquired wireless data and task features, we build multiple models to evaluate users different task engagements. While we can distinct between busy and relax time frames with an accuracy of just above 75% and 68% of accuracy between the transitions of the two, predicting the actual level of

task engagement proved to be too challenging for our system. Still, for certain task types, such as the *number comparison* task, we could differentiate between two task difficulties with the accuracy of 65%. Using deep neural network approach with a raw signal input (without extracting features) has led to better results at some problem definitions. Apparently, while extracting heartbeat and respiratory features manually, such as time domain and frequency domain features, we did not manage to cover all of them and deep learning has the advantage of finding these hidden features with a price of higher computational power consumption.

Being optimistic at first, we came to a conclusion that assessing one's cognitive load using only wireless signal reflections is a challenging task. Being present while conducting a user study have led to interesting, yet not surprising remarks. All users were told to make small movements (e.g. no excessive limb motion), due to system sensitivity. Still, this was difficult for some of the volunteers, as people have different habits, that they are not even aware of (e.g. touching face with one of the hands while solving a task). Still, some of the results of Wi-Mind could already be beneficial. E.g. we noticed that we can predict when someone is starting of finishing some task. This could be either doing homework, studying, reading some article, etc. Applying Wi-Mind could in theory be able to detect start in cognitive engagement and turn off all notifications until you finish your task engagement. This does not only apply to computer related work, but also on other stationary tasks, e.g. reading newspapers, watching television, listening to podcasts, etc. Another application domain could be improving user engagement in studying, especially with interactive educational materials. Here, upon Wi-Mind detecting reduced engagement, the application could pop up an interesting instructional video or a short quiz. Cognitive load inference also raises ethical issues. A video game, for example, could be adapted to manipulate a player's arousal to their inferred cognitive load, potentially leading to gaming addiction [59]. Similar usage could be applied in social media, e.g. showing different posts to keep person engaged could

waste more time from an individual user.

Finally, we should be aware that numerous factors "sit" between the objective complexity of the task and our physiological reaction. These include our motivation to solve a task, our cognitive capacities, as well as our personality and skills. Furthermore, the cognitive load proxied through our physiological reaction is a complex concept that reflects the task's inherent complexity (i.e. intrinsic load), the complexity of the task's representation (i.e. extraneous load), and the complexity of constructing the schema of the task (i.e. germane load) [60]. The plethora of concepts involved and multiple levels of indirection make cognitive load inference a very challenging problem. In this thesis we present a pioneering approach towards wireless cognitive load inference, yet aware of the work's limitations, we call for further investigation of this exciting research field.

Appendix A

Demographic data form and study consent

Starost: ____

Spol: M Ž

Moja dominantna roka je: LEVA DESNA

Stopnja izobrazbe: I. II. III. IV. V. VI/1. VI/2. VII. VIII/1. VIII/2

nedokončana OŠ	I.	
OŠ	II.	
nižje poklicno izobraževanje (2 letno)	III.	
srednje poklicno izobraževanje (3 letno)	IV.	
gimnazijsko, srednje poklicno -tehniško izobraževanje, srednje tehniško oz. drugo strokovno izobraževanje	V.	
višješolski program (do 1994), višješolski strokovni program	VI/1.	
specializacija po višješolskem programu, visokošolski strokovni programi	VI/2.	visokošolski strokovni in univerzitetni program (1. bol. st)
specializacija po visokošolskem strokovnem programu, univerzitetni program	VII.	magisterij stroke (2. bol. st.)
specializacija po univerzitetnem programu, magisterij znanosti	VIII/1.	
doktorat znanosti	VIII/2.	doktorat znanosti (3. bol. st.)

Izpolni raziskovalec:

ID: _____

Sporazum o zavestnem in prostovoljnem sodelovanju v raziskavi
S svojim podpisom potrjujem naslednje:

- Vsebina študije mi je bila razložena v celoti in sem jo razumel.
- Na vsa moja dodatna vprašanja sem prejel odgovore, ki so v celoti zadovoljivi.
- Strinjam se s sodelovanjem v študiji.
- Zavedam se, da je sodelovanje v študiji prostovoljno.
- Seznanjen sem, da lahko v kateremkoli trenutku tekom izvajanja eskperimenta prekličem sodelovanje.

Izjava o varovanju podatkov
Strinjam se, da se zbrani podatki uporabijo v znanstvene namene. Za vse zbrane podatke bo zagotovljena popolna anonimizacija. Prav tako se strinjam, da se zbrani podatki lahko uporabijo za to študijo in za naslednje študije ter publikacije. Anonimizirani podatki se lahko uporabijo s strani raziskovalcev in sodelavcev Fakultete za računalništvo in informatiko, Univerze v Ljubljani, ter zunanjih raziskovalcev, po dovoljenju Fakultete za računalništvo in informatiko.

Umik sporazuma o uporabi zbranih podatkov
Seznanjen sem, da lahko kadarkoli in brez razloga umaknem sporazum za uporabo podatkov. V primeru umika sporazuma se strinjam, da se podatki shranijo v kontrolne namene. Prav tako imam pravico zahtevati izbris podatkov. Zavedam se tudi, da v primeru izvedene anonimizacije nad podatki, zahtevan izbris ni mogoč.

Ljubljana, dne _____

Podpis udeleženca

_____

Bibliography

[1] S. Leroy, Why is it so hard to do my work? The challenge of attention residue when switching between work tasks, Organizational Behavior and Human Decision Processes 109 (2) (2009) 168–181.

[2] G. Mark, D. Gudith, U. Klocke, The cost of interrupted work: more speed and stress, in: Proceedings of the SIGCHI conference on Human Factors in Computing Systems, ACM, Florence, Italy, 2008, pp. 107–110.

[3] S. G. Hart, L. E. Staveland, Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research, in: Advances in psychology, Vol. 52, Elsevier, 1988, pp. 139–183.

[4] M. Shakouri, L. H. Ikuma, F. Aghazadeh, I. Nahmens, Analysis of the sensitivity of heart rate variability and subjective workload measures in a driving simulator: the case of highway work zones, International Journal of Industrial Ergonomics 66 (2018) 136–145.

[5] M. Grassmann, E. Vlemincx, A. von Leupoldt, J. M. Mittelstädt, O. Van den Bergh, Respiratory changes in response to cognitive load: A systematic review, Neural plasticity 2016.

[6] M. Haak, S. Bos, S. Panic, L. Rothkrantz, Detecting stress using eye blinks and brain activity from EEG signals, Proceeding of the 1st driver car interaction and interface (DCII 2008) (2009) 35–60.

[7] Y. Abdelrahman, E. Velloso, T. Dingler, A. Schmidt, F. Vetere, Cognitive heat: exploring the usage of thermal imaging to unobtrusively estimate cognitive load, Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 1 (3) (2017) 33.

[8] D. McDuff, S. Gontarek, R. Picard, Remote measurement of cognitive stress via heart rate variability, in: 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, Chicago, Illinois, USA, 2014.

[9] F. Adib, H. Mao, Z. Kabelac, D. Katabi, R. C. Miller, Smart homes that monitor breathing and heart rate, in: Proceedings of the 33rd annual ACM conference on human factors in computing systems, ACM, Seoul, Republic of Korea, 2015, pp. 837–846.

[10] X. Wang, C. Yang, S. Mao, TensorBeat: Tensor decomposition for monitoring multiperson breathing beats with commodity WiFi, ACM Transactions on Intelligent Systems and Technology (TIST) 9 (1) (2017) 8.

[11] R. Ravichandran, E. Saba, K.-Y. Chen, M. Goel, S. Gupta, S. N. Patel, WiBreathe: Estimating respiration rate using wireless signals in natural settings in the home, in: IEEE International Conference on Pervasive Computing and Communications (PerCom), IEEE, St. Louis, MO, USA, 2015, pp. 131–139.

[12] X. Huang, L. Sun, T. Tian, Z. Huang, E. Clancy, Real-time non-contact infant respiratory monitoring using UWB radar, in: IEEE 16th International Conference on Communication Technology (ICCT), IEEE, Hangzhou, China, 2015, pp. 493–496.

[13] A. Droitcour, V. Lubecke, J. Lin, O. Boric-Lubecke, A microwave radio for doppler radar sensing of vital signs, in: IEEE MTT-S International Microwave Symposium Digest, Vol. 1, IEEE, Phoenix, AZ, USA, 2001, pp. 175–178.

[14] O. B. Lubecke, P.-W. Ong, V. Lubecke, 10 GHz Doppler radar sensing of respiration and heart movement, in: Proceedings of the IEEE 28th Annual Northeast Bioengineering Conference, IEEE, Philadelphia, PA, USA, 2002, pp. 55–56.

[15] R. Nandakumar, S. Gollakota, N. Watson, Contactless sleep apnea detection on smartphones, in: Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services, ACM, Florence, Italy, 2015, pp. 45–57.

[16] E. Haapalainen, S. Kim, J. F. Forlizzi, A. K. Dey, Psycho-physiological measures for assessing cognitive load, in: ACM UbiComp'10, Copenhagen, Denmark, 2010.

[17] J. Demšar, T. Curk, A. Erjavec, Č. Gorup, T. Hočevar, M. Milutinovič, M. Možina, M. Polajnar, M. Toplak, A. Starič, et al., Orange: data mining toolbox in Python, The Journal of Machine Learning Research 14 (1) (2013) 2349–2353.

[18] Keras: The Python Deep Learning library, `https://keras.io`, [Online; accessed 07-September-2018].

[19] M. Weiser, The Computer for the 21st Century, Scientific American 265 (3) (1991) 94–104.

[20] A. Mehrotra, M. Musolesi, R. Hendley, V. Pejovic, Designing Content-driven Intelligent Notification Mechanisms for Mobile Applications, in: UbiComp'15, ACM, Osaka, Japan, 2015, pp. 813–824.

[21] A. Mehrotra, V. Pejovic, J. Vermeulen, R. Hendley, M. Musolesi, My phone and me: Understanding people's receptivity to mobile notifications, in: ACM CHI'16, San Jose, CA, USA, 2016.

[22] V. Pejovic, A. Mehrotra, M. Musolesi, Investigating The Role of Task Engagement in Mobile Interruptibility, in: Smarttention, Please! Intel-

ligent Attention Management on Mobile Devices Workshop (with MobileHCI'15), Copenhagen, Denmark, 2015.

[23] C. Anderson, I. Hübener, A.-K. Seipp, S. Ohly, K. David, V. Pejovic, A survey of attention management systems in ubiquitous computing environments, arXiv preprint arXiv:1806.06771.

[24] T. K. Fredericks, S. D. Choi, J. Hart, S. E. Butt, A. Mital, An investigation of myocardial aerobic capacity as a measure of both physical and cognitive workloads, International Journal of Industrial Ergonomics 35 (12) (2005) 1097–1107.

[25] M. ElKomy, Y. Abdelrahman, M. Funk, T. Dingler, A. Schmidt, S. Abdennadher, ABBAS: an adaptive bio-sensors based assistive system, in: Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems, ACM, Denver, Colorado, USA, 2017, pp. 2543–2550.

[26] Y. Shi, N. Ruiz, R. Taib, E. Choi, F. Chen, Galvanic skin response (GSR) as an index of cognitive load, in: CHI'07 extended abstracts on Human factors in computing systems, ACM, San Jose, CA, USA, 2007, pp. 2651–2656.

[27] Z. Wang, J. Yan, H. Aghajan, A framework of personal assistant for computer users by analyzing video stream, in: Proceedings of the 4th Workshop on Eye Gaze in Intelligent Human Machine Interaction, ACM, Santa Monica, CA, USA, 2012, p. 14.

[28] T. Matkovič, V. Pejovic, Wi-mind: Wireless mental effort inference, in: ACM UbitTention workshop'18, Singapore, Singapore, 2018.

[29] K. Whitenton, Minimize Cognitive Load to Maximize Usability, `https://www.nngroup.com/articles/minimize-cognitive-load/`, [Online; accessed 23-September-2018] (2018).

[30] J. P. Borst, N. A. Taatgen, H. Van Rijn, The problem state: A cognitive bottleneck in multitasking., Journal of Experimental Psychology: Learning, memory, and cognition 36 (2) (2010) 363.

[31] Y. Miyata, D. A. Norman, Psychological issues in support of multiple activities, User centered system design: New perspectives on human-computer interaction (1986) 265–284.

[32] M. T. Ullman, Contributions of memory circuits to language: The declarative/procedural model, Cognition 92 (1-2) (2004) 231–270.

[33] J. P. Borst, N. A. Taatgen, H. van Rijn, What makes interruptions disruptive?: A process-model account of the effects of the problem state bottleneck on task interruption and resumption, in: Proceedings of the 33rd annual ACM conference on human factors in computing systems, ACM, Seoul, Republic of Korea, 2015, pp. 2971–2980.

[34] F. Paas, J. E. Tuovinen, H. Tabbers, P. W. Van Gerven, Cognitive load measurement as a means to advance cognitive load theory, Educational psychologist 38 (1) (2003) 63–71.

[35] G. B. Reid, T. E. Nygren, The subjective workload assessment technique: A scaling procedure for measuring mental workload, in: Advances in psychology, Vol. 52, Elsevier, 1988, pp. 185–218.

[36] M. Züger, T. Fritz, Interruptibility of software developers and its prediction using psycho-physiological sensors, in: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, ACM, Seoul, Republic of Korea, 2015, pp. 2981–2990.

[37] M. Gjoreski, H. Gjoreski, M. Luštrek, M. Gams, Continuous stress detection using a wrist device: in laboratory and real life, in: Mental Health workshop with ACM UbiComp'16, Heidelberg, Germany, 2016.

[38] G. Balakrishnan, F. Durand, J. Guttag, Detecting pulse from head motions in video, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, Oregon, 2013.

[39] M. Zhao, F. Adib, D. Katabi, Emotion recognition using wireless signals, in: Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking, ACM, New York City, New York, 2016, pp. 95–108.

[40] G. Urh, V. Pejovic, Taskyapp: Inferring task engagement via smartphone sensing, in: ACM UbitTention workshop'16, Heidelberg, Germany, 2016.

[41] T. Matkovič, Wi-Mind - Github repository, `https://github.com/matkovic/Wi-Mind`, [Online; accessed 14-October-2018] (2018).

[42] S. Wunsch, gr-radar - GNU Radio Radar Toolbox, `https://github.com/kit-cel/gr-radar`, [Online; accessed 20-July-2018] (2018).

[43] GNU Radio, `https://www.gnuradio.org`, [Online; accessed 26-July-2018] (2018).

[44] F. Adib, Z. Kabelac, D. Katabi, R. C. Miller, 3D Tracking via Body Radio Reflections, in: NSDI, Vol. 14, Seattle, WA, 2014, pp. 317–329.

[45] H. Forstén, Heartbeat detection with radar, `http://hforsten.com/heartbeat-detection-with-radar.html`, [Online; accessed 23-September-2018] (2018).

[46] New to SDR?, `http://luaradio.io/new-to-sdr.html`, [Online; accessed 04-September-2018] (2018).

[47] R. W. Stewart, K. W. Barlee, D. S. Atkinson, L. H. Crockett, Software defined radio using MATLAB & Simulink and the RTL-SDR, Strathclyde Academic Media, 2015.

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

[49] J. Wijsman, B. Grundlehner, H. Liu, J. Penders, H. Hermens, Wearable physiological sensors reflect mental stress state in office-like situations, in: 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII), IEEE, Geneva, Switzerland, 2013, pp. 600–605.

[50] C. Maaoui, A. Pruski, F. Abdat, Emotion recognition for human-machine communication, in: IEEE/RSJ International Conference on Intelligent Robots and Systems, IEEE, Nice, France, 2008, pp. 1210–1215.

[51] A. Alberdi, A. Aztiria, A. Basarab, Towards an automatic early stress recognition system for office environments based on multimodal measurements: A review, Journal of biomedical informatics 59 (2016) 49–75.

[52] A. Kaklauskas, E. K. Zavadskas, M. Seniut, G. Dzemyda, V. Stankevic, C. Simkevičius, T. Stankevic, R. Paliskiene, A. Matuliauskaite, S. Kildiene, et al., Web-based biometric computer mouse advisory system to analyze a user's emotions and work productivity, Engineering Applications of Artificial Intelligence 24 (6) (2011) 928–945.

[53] P. Jönsson, Respiratory sinus arrhythmia as a function of state anxiety in healthy individuals, International journal of psychophysiology 63 (1) (2007) 48–54.

[54] M. J. Reed, C. Robertson, P. Addison, Heart rate variability measurements and the prediction of ventricular arrhythmias, Qjm 98 (2) (2005) 87–95.

[55] B. Cinaz, B. Arnrich, R. Marca, G. Tröster, Monitoring of mental workload levels during an everyday life office-work scenario, Personal and ubiquitous computing 17 (2) (2013) 229–239.

[56] A. Mital, M. Govindaraju, Is it possible to have a single measure for all work?, International journal of industrial engineering-theory applications and practice 6 (3) (1999) 190–195.

[57] Microsoft Band, `https://www.microsoft.com/en-us/band`, [Online; accessed 14-September-2018].

[58] P. Nickel, F. Nachreiner, Sensitivity and diagnosticity of the 0.1-Hz component of heart rate variability as an indicator of mental workload, Human Factors 45 (4) (2003) 575–590.

[59] A. Parnandi, Y. Son, R. Gutierrez-Osuna, A control-theoretic approach to adaptive physiological games, in: 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, IEEE, Geneva, Switzerland, 2013, pp. 7–12.

[60] J. Sweller, J. J. Van Merrienboer, F. G. Paas, Cognitive architecture and instructional design, Educational psychology review 10 (3) (1998) 251–296.