Univerza v Ljubljani
Fakulteta za računalništvo in informatiko

Gašper Slapničar

Trajno napovedovanje krvnega tlaka iz signala PPG

MAGISTRSKO DELO

MAGISTRSKI PROGRAM DRUGE STOPNJE
RAČUNALNIŠTVO IN INFORMATIKA

Mentor: izr. prof. dr. Matjaž Kukar

Somentor: viš. zn. sod. dr. Mitja Luštrek

Ljubljana, 2018

University of Ljubljana
Faculty of Computer and Information Science

Gašper Slapničar

Continuous blood pressure estimation from PPG signal

MASTER'S THESIS

THE 2ND CYCLE MASTER'S STUDY PROGRAMME
COMPUTER AND INFORMATION SCIENCE

Supervisor: izr. prof. dr. Matjaž Kukar

Co-supervisor: viš. zn. sod. dr. Mitja Luštrek

Ljubljana, 2018

Povzetek

Naslov: Trajno napovedovanje krvnega tlaka iz signala PPG

Krvni tlak je pomemben pokazatelj hipertenzije. Razvili smo sistem, ki krvni tlak ocenjuje iz fotopletizmograma (PPG), kakršen je že vgrajen v večino modernih senzorskih zapestnic. Zaradi šuma in motenj, ki se v signalu PPG pojavijo kot posledica uporabe zapestnice, smo razvili metodo čiščenja in segmentiranja signala PPG na cikle. Nato smo izračunali množico značilk, ki smo jih uporabili v regresijskih modelih. Sistem smo izboljšali z uporabo algoritma RReliefF za izbor najboljših značilk in z uporabo dela podatkov vsake osebe za učenje personaliziranih napovednih modelov. Sistem smo vrednotili na dveh podatkovnih množicah, eni iz kliničnega okolja in drugi zbrani med rutinskimi dnevnimi aktivnostmi posameznikov. V poizkusu, kjer model vsakič naučimo na vseh osebah razen eni in ga nato testiramo na izpuščeni osebi, smo z uporabo klinične množice (podatkovna baza MIMIC) dosegli najnižjo povprečno absolutno napako (MAE) 5,61 mmHg za sistolični in 3,82 mmHg za diastolični krvni tlak, oboje pri največji stopnji personalizacije. Za množico, zbrano med rutinskimi dnevnimi aktivnostmi, smo dosegli najnižjo MAE 8.40 mmHg za sistolični in 4.20 mmHg za diastolični krvni tlak, ponovno pri največji stopnji personalizacije. Najbolje sta se obnesla algoritma globoka regresija in “naključni gozd”. Rezultati skoraj dosegajo zahteve dveh glavnih standardov za ocenjevanje krvenga tlaka.

Ključne besede

krvni tlak, fotopletizmografija, strojno učenje, regresija, obdelava signalov

Abstract

Title: Continuous blood pressure estimation from PPG signal

Blood pressure (BP) is an indicator of hypertension. We developed a system in which photoplethysmogram (PPG), which is commonly integrated in modern wearables, is used to continuously estimate BP. A preprocessing module was developed and used for cleaning the PPG signal of noise and artefacts, and segmenting it into cycles. A set of features describing the PPG signal was then computed to be used in regression models. The RReliefF algorithm was used to select a subset of relevant features and personalization of the models was considered to further improve the performance of the models. The approach was validated using two distinct datasets, one from a hospital environment, and the other collected during every-day activities. Using the clinical dataset (MIMIC database), the best achieved mean absolute errors (MAE) in a leave-one-subject-out (LOSO) experiment were 5.61 mmHg for systolic and 3.82 mmHg for diastolic BP, at maximum personalization. For everyday-life dataset, the lowest errors were 8.40 mmHg for systolic and 4.20 mmHg for diastolic BP. Deep learning regression and Random Forest algorithm achieved the best results. Our results borderline meet the requirements of the two most well-established standards for BP estimation devices.

Keywords

blood pressure, photoplethysmography, machine learning, regression, signal processing

Acknowledgments

I would first like to thank my supervisor dr. Matjaž Kukar for his insights, patience and guidance during the creation of this thesis. A special thanks goes to dr. Mitja Luštrek, who has offered me the opportunity to work with him, and has set an example of excellence as a researcher and mentor.

I would also like to thank my fellow graduate students, who have helped me throughout my studies. I am also very grateful to everyone at the Department of Intelligent Systems at Jožef Stefan Institute, who have shared their knowledge, experience, and ideas with me.

Finally, I would like to thank my family, who have shown immense patience and support during my studies.

Gašper Slapničar, 2018

Contents

List of used acronmys

acronym	**meaning**
BP	blood pressure
PPG	photoplethysmogram
ECG	electrocardiogram
ABP	arterial blood pressure
SBP	systolic blood pressure
DBP	diastolic blood pressure
PD	photodetector
LED	light-emitting diode
PAT	pulse arrival time
PTT	pulse transit time
IBI	interbeat interval
SACF	sample autocorrelation function
SQI	signal quality index
DTW	dynamic time warping
FFT	fast fourier transform
MAE	mean absolute error
ANN	artificial neural network
AAMI	Advancement of Medical Instrumentation
BHS	British Hypertension Society
WHO	World Health Organization

Razširjeni povzetek

I Uvod

Bolezni srca in ožilja so bile leta 2015 skupno gledano najpogostejši bolezenski vzrok smrti [1]. Povišan krvni tlak je glavni simptom, ki nakazuje tovrstno bolezensko stanje, zato bi ljudje morali redno spremljati svoj krvni tlak. Redno spremljanje krvnega tlaka je sploh ključno pri bolnikih, ki že trpijo za takšnimi boleznimi, in tistih, ki imajo povečano tveganje za pojav le-teh.

Tradicionalna metoda za ocenjevanje krvnega tlaka z uporabo napihljive manšete ni uporabna za spremljanje krvnega tlaka med fizično aktivnostjo ali spanjem, saj sama naprava resno omejuje bolnikovo gibanje. Poleg tega protokol merjenja zahteva specifičen položaj manšete v višini srca, kar je dodatna omejitev [2].

Postopek merjenja z manšeto tipično izvaja osebje v bolnišnicah, kar lahko povzroči občutek nelagodja in stresa pri bolniku in lahko posledično vpliva na dejansko vrednost krvnega tlaka [3]. Poleg tega kompleksnost merjenja povzroča odklonilen odnos bolnikov, ki si zato tlaka ne merijo tako pogosto, kot bi morali [4].

Zaradi opisanih dejavnikov bi bilo smiselno razviti sistem, ki bi omogočal natančno in stalno spremljanje krvnega tlaka na neinvaziven način, torej brez potrebe po napihljivi manšeti. To bi poenostavilo proces merjenja in posledično verjetno zmanjšalo odklonilnost bolnikov do le-tega. Poleg tega bi omogočilo merjenje v primerih, ko uporaba manšete ni možna (npr. aktivnost, spanje). Razširjene možnosti uporabe in manjša odklonilnost bolnikov bi povečala njihovo osveščenost o trenutnem zdravstvenem stanju.

Dober kandidat za razvoj takšnega sistema so senzorske zapestnice z vgrajenim senzorjem za merjenje fotopletizmograma (PPG), ki se pogosto uporablja za ocenjevanje srčnega utripa. PPG je osnovan na presvetljevanju tkiva (v zapestnicah tipično z zeleno svetlobo [5], ki je bolj odporna proti motnjam) in merjenju sprememb absorbcije svetlobe. Z vsakim srčnim utripom srce potisne kri proti robnim točkam v telesu. Vsak utrip srca se kaže v pripadajočem ciklu signala PPG, ki prikazuje tudi trenutno količino krvi v tkivu. Sprememba količine krvi v tkivu vpliva na krvni tlak (npr. ko je v žili več krvi je pritisk na stene žil večji) in zato je ta signal uporaben za ocenjevanje krvnega tlaka [6, 7].

II Sorodno delo

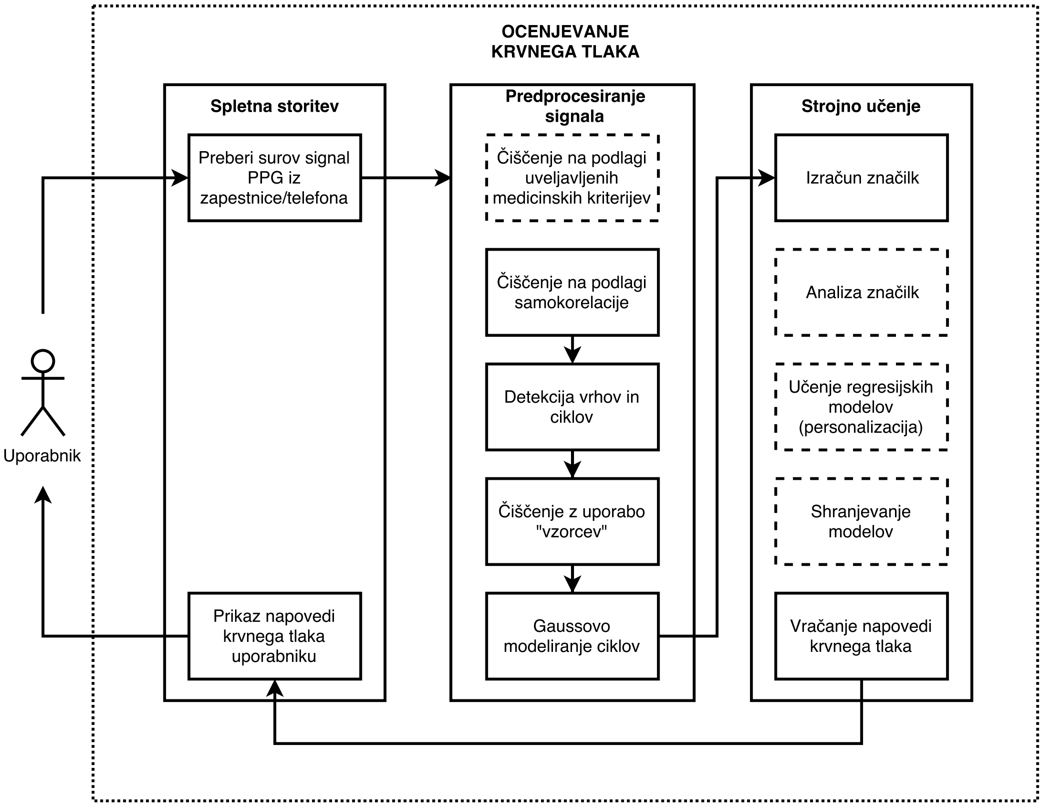
Sorodno delo na tem področju večinoma obravnava enega izmed dveh pristopov, ki sta se uveljavila za ocenjevanje krvnega tlaka iz signala PPG.

Prvi je osnovan na času, ki je potreben, da se kri prenese od srca do neke periferne točke v telesu v okviru enega srčnega utripa. Ta čas se imenuje "pulse transit time (PTT)", povezava sprememb tega časa s spremembami krvnega tlaka pa je dokazana in uveljavljena [8, 9, 10, 11]. Ta čas je krajši v primeru, ko so stene žil bolj čvrste, kar pospeši pretok krvi in implicira večji pritisk krvi na stene žil. Slabost tega pristopa je potreba po uporabi dveh senzorjev, tipično senzorja za elektrokardiogram (EKG) in senzorja za PPG.

Drugi pristop, ki je v zadnjih letih raziskovalno vse bolj zanimiv, stremi k odpravi dveh senzorjev in poskuša ocenjevati krvni tlak samo iz signala PPG. Prvi večji problem tega pristopa je stik med senzorjem zapestnice in kožo, ki je med aktivnostjo zaradi premikanja roke, zapestja in zapestnice pogosto moten. Drugi problem je manjko jasne klinično potrjene povezave med PPG-jem in krvnim tlakom. Ob določenih predpostavkah se korelacija med PPG-jem in krvnim tlakom kaže v sorodnih delih, vendar je še vedno predmet aktivnih raziskav in te magistrske naloge [12, 13, 14, 15, 16].

III Metodologija

Sistem sestoji iz dveh modulov, modula za predprocesiranje signala in modula za strojno učenje. Prvi je odgovoren za segmentacijo signala PPG na cikle in odstranjevanje ciklov z morfološkimi spremembami zaradi šuma. Drugi skrbi za izračun značilk, ki opišejo PPG signal. Poleg tega izbere podmnožico relevantnih značilk in jih uporabi v regresijskih algoritmih strojnega učenja z namenom učenja napovednih modelov za sistolični in diastolični krvni tlak. Ta modula dopolnjuje spletna storitev, ki omogoča interakcijo med zapestnico oz. pametnim telefonom in omenjenima moduloma. Opisana arhitektura je prikazana na sliki 1.


OCENJEVANJE KRVNEGA TLAKA
Spletna storitev
Preberi surov signal PPG iz zapestnice/telefona
Prikaz napovedi krvnega tlaka uporabniku
Predprocesiranje signala
Čiščenje na podlagi uveljavljenih medicinskih kriterijev
Čiščenje na podlagi samokorelacije
Detekcija vrhov in ciklov
Čiščenje z uporabo "vzorcev"
Gaussovo modeliranje ciklov
Strojno učenje
Izračun značilk
Analiza značilk
Učenje regresijskih modelov (personalizacija)
Shranjevanje modelov
Vračanje napovedi krvnega tlaka
Uporabnik

Slika 1: Predlagana arhitektura sistema za ocenjevanje krvnega tlaka. Črtkani pravokotniki predstavljajo korake, ki so potrebni le enkrat ali periodično (učenje prvega modela ali personalizacija).

V prvi fazi predprocesiranja se odstranijo 5-sekundni segmenti signala PPG, ki ustrezajo skrajnim ali pa klinično nemogočim vrednostim krvnega tlaka (npr. krvni tlak 0 mmHg ali nenadna sprememba v 5 sekundah za več kot 50 mmHg). Meje za razumne vrednosti krvnega tlaka so bile povzete po sorodnem delu [16] in nato rahlo prilagojene glede na opažanja v naših podatkih.

V drugi fazi se odstranijo 5-sekundni segmenti PPG signala, ki imajo zelo nizko samokorelacijo. Zaradi periodične narave PPG signala in podobnosti ciklov je pričakovano, da bo samokorelacija segmenta, ki vsebuje nekaj ciklov, postala ob zamiku za trajanje enega cikla zelo visoka. V primeru, da kratek

PPG segment vsebuje veliko šuma, samokorelacija nikoli ne bo zelo visoka, ne glede na zamik.

V tretji fazi se signal PPG segmentira na cikle, ki ustrezajo srčnim utripom. Uporabili smo algoritem, ki so ga prvi predlagali Lazaro et al. [17] in je specializiran za zaznavo sistoličnih vrhov PPG signala. Algoritem je osnovan na filtru, ki zazna nenadne strme vzpone v signalu, ki tipično nakazujejo bližino sistoličnega vrha. Ta filter je dopolnjen z algoritmom za dinamično določanje amplitude, nad katero mora biti vrh, da ga algoritem sprejme kot pravi vrh in ne kot šum ali diastolični vrh. Ko so vrhovi zaznani, na preprost način zaznamo še začetne in končne točke ciklov, ki ustrezajo najnižjim dolinam med vrhovi.

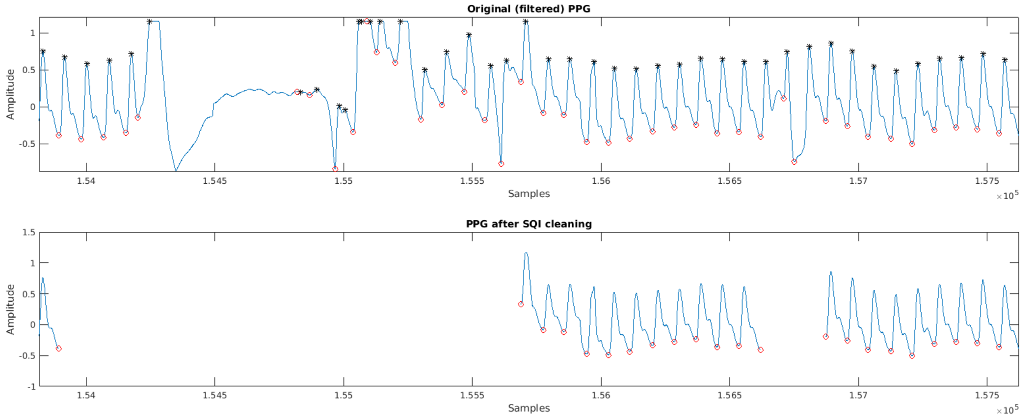
Ko so cikli zaznani, se v četrti fazi ustvarijo t.i. “vzorci” (angl. templates – T) ciklov. Najprej se s pomočjo samokorelacijske analize določi najbolj verjetna dolžina cikla v 30-sekundnem PPG segmentu, ki jo označimo z L. Nato se od vsake začetne točke cikla vzame L vzorcev signala in se izračuna T kot povprečje vrednosti signala vseh ciklov ob enakem času. Ko je T izračunan se izračuna podobnost vsakega cikla s T-jem glede na tri mere kakovosti cikla (angl. signal quality indices – SQIs):

1. *SQI1* – direktna linearna korelacija z uporabo Pearsonovega koeficienta, kjer se vzame L vzorcev od začetne točke vsakega cikla,

2. *SQI2* – direktna linearna korelacija, ponovno z uporabo Pearsonovega koeficienta, vendar je tokrat vsak cikel prevzorčen na dolžino L,

3. *SQI3* – korelacija med ciklom in T-jem z uporabo tehnike “Dynamic Time Warping (DTW)”.

Zahtevane vrednosti za vsak SQI so bile povzete po sorodnem delu [18]. Tisti cikli, ki niso dosegali zahtevanih vrednosti, se zavržejo. Rezultat čiščenja je prikazan na sliki 2.



Slika 2: Zgornja slika prikazuje PPG signal z očitnimi artefakti, medtem ko spodnja prikazuje rezultat po predprocesiranju, ki uspešno odstrani artefakte.

Ker klinično pričakovana oblika PPG cikla sestoji iz dveh zaporednih vrhov (sistoličnega in diastoličnega), se posamezen cikel da modelirati z vsoto dveh Gaussovih funkcij [15]. Z upoštevanjem določenih omejitev (npr. sistolični vrh mora biti pred diastoličnim, sistolični vrh mora biti višji kot diastolični itd.) lahko najdemo le takšne cikle, ki se dobro prilegajo klinično pričakovani obliki z dvema vrhovoma. Takšna oblika je pomembna za izračun značilk, ki upoštevajo diastolični vrh.

Ko smo uspešno pridobili kvalitetne cikle, smo iz njih izračunali množico značilk. Osnovno množico smo povzeli po sorodnem delu in jo dopolnili z lastnimi značilkami, kot so ploščine ter značilke iz frekvenčnega prostora in iz analize kompleksnosti signala [12, 13, 14, 15, 16]. Nato smo izmed vseh značilk izbrali najboljše z uporabo algoritma RReliefF [19].

Izbrane značilke smo uporabili v regresijskih algoritmih za učenje napovednih modelov. Uporabili smo širok nabor algoritmov iz klasične regresije (regresijska drevesa, modelna drevesa, naključni gozd itd.), simbolično regresijo z genetskim algoritmom [20] ter regresijo globokega učenja [21].

IV Podatki

Uporabili smo dve podatkovni množici, eno iz kliničnega okolja in drugo zbrano med rutinskimi dnevnimi aktivnostmi posameznikov.

Prvo smo pridobili iz prosto dostopne podatkovne baze MIMIC [22], v kateri so fiziološki signali bolnikov, ki so bili zbrani v bolnišnici. Uporabili smo podatke vseh bolnikov, ki so imeli na voljo PPG in krvni tlak.

Druga podatkovna množica je bila zbrana med dnevnimi aktivnostmi zaposlenih na Institutu Jožef Stefan. PPG se je meril z uporabo senzorske zapestnice Empatica E4, medtem ko se je referenčni krvni tlak periodično meril z uporabo digitalnega merilnika krvnega tlaka Omron.

Prva množica je bila mnogo večja in signal je bil v splošnem bolj kvaliteten, medtem ko je bila druga množica manjša in bolj šumna, kar je posledica merjenja z zapestnico v nenadzorovanem okolju.

V Poizkusi in vrednotenje

Uporabili smo tri oz. štiri poizkuse za vrednotenje napovednih modelov, in sicer enega za simbolično regresijo ter tri za ostale regresijske algoritme. Prvi in drugi poizkus sta učno in testno množico ustvarila iz primerov enega osebka s preprosto delitvijo 66% - 34%, tretji je uporabil prečno preverjanje (angl. cross-validation – CV), zadnji pa je vedno izpustil en osebek za testiranje in se učil na vseh ostalih (angl. Leave-one-subject-out – LOSO).

V vseh poizkusih smo za mero uspešnosti uporabili povprečno absolutno napako (angl. Mean Absolute Error – MAE), ki kaznuje tako pozitivna kot negativna odstopanja napovedi od dejanskih vrednosti s tem, da vzame absolutno vrednost odstopanja.

Simbolična regresija se že v bolj preprostem poizkusu z delitvijo podatkov na učno (66%) in testno (34%) množico ni izkazala, zato smo jo za ostale poizkuse opustili.

Med ostalimi je poizkusalno najbolj zahteven in zanimiv poizkus LOSO, saj poskuša ustvariti splošen model in ga vedno testira ne nekem osebku ki ni bil nikoli uporabljen med učenjem. Sprva je ta poizkus v primerjavi z ostalimi dosegal slabše rezultate, vendar smo modele izboljšali s personalizacijo. To smo dosegli tako, da smo v učno množico dodajali manjše število primerov izpuščenega testnega osebka (10% – 50%) in tako splošen model rahlo prilagodili temu osebku. Takšen poizkus simulira situacijo v kateri si osebek nekajkrat sam pomeri krvni tlak z natančnim merilnikom in ga nato vnese v sistem, ki sproti meri tudi njegov PPG. Sistem se tako prilagodi dotičnemu osebku.

Z uporabo klinične podatkovne množice se je v poizkusu LOSO najbolje obnesla regresija globokega učenja, ki je dosegla najnižjo MAE 5.61 mmHg za sistolični in 3.82 za diastolični krvni tlak, oboje pri največji stopnji personalizacije, kar je prikazano v tabeli 1.

Najboljše v poizkusu 2				
Algoritem	**MAE$_{\mathbf{SBP}}$**	**STD$_{\mathbf{SBP}}$**	**MAE$_{\mathbf{DBP}}$**	**STD$_{\mathbf{DBP}}$**
Random Forest	6.23	6.92	4.53	3.62
Najboljše v poizkusu 3				
Algoritem	**MAE$_{\mathbf{SBP}}$**	**STD$_{\mathbf{SBP}}$**	**MAE$_{\mathbf{DBP}}$**	**STD$_{\mathbf{DBP}}$**
Random Forest	7.83	7.47	3.84	3.63
Najboljše v poizkusu 4				
	MAE$_{\mathbf{SBP}}$	**STD$_{\mathbf{SBP}}$**	**MAE$_{\mathbf{DBP}}$**	**STD$_{\mathbf{DBP}}$**
Algoritem	0% / 50%	0% / 50%	0% / 50%	0% / 50%
Random Forest	18.66 / 6.32	14.16 / 6.02	10.07 / 4.32	6.31 / 3.37
Deep learning	17.12 / 5.61	15.22 / 6.12	9.82 / 3.82	5.42 / 2.91

Tabela 1: Primerjava najboljših algoritmov v poizkusih 2, 3 in 4 za klinično podatkovno množico. Podane so povprečne absolutne napake in standardni odkloni. Vsi rezultati so v mmHg. Odstotki pomenijo količino podatkov uporabljenih za personalizacijo.

Za podatkovno množico, zbrano med dnevnimi aktivnostmi uporabnikov, regresija globokega učenja ni bila smiselna, saj je bilo v trenutni fazi zbiranja na voljo premalo podatkov. Posledično se v poizkusu LOSO najbolj izkazal algoritem Random Forest, ki je dosegel najnižjo MAE 8.40 mmHg za sistolični in 4.20 mmHg za diastolični krvni tlak, ponovno pri največji stopnji personalizacije, kar je prikazano v tabeli 2.

Najboljše v poizkusu 3				
Algoritem	$\mathbf{MAE_{SBP}}$	$\mathbf{STD_{SBP}}$	$\mathbf{MAE_{DBP}}$	$\mathbf{STD_{DBP}}$
Random Forest	8.92	8.49	4.27	3.99
Najboljše v poizkusu 4				
	$\mathbf{MAE_{SBP}}$	$\mathbf{STD_{SBP}}$	$\mathbf{MAE_{DBP}}$	$\mathbf{STD_{DBP}}$
Algoritem	0% / 50%	0% / 50%	0% / 50%	0% / 50%
Random Forest	12.81 / 8.40	11.03 / 7.53	7.19 / 4.20	5.29 / 3.18

Tabela 2: Primerjava najboljših algoritmov v poizkusih 3 in 4 za podatkovno množico, zbrano med dnevnimi aktivnostmi. Podane so povprečne absolutne napake in standardni odkloni. Vsi rezultati so v mmHg. Odstotki pomenijo količino podatkov, uporabljenih za personalizacijo.

VI Zaključek

Primerjava s sorodnim delom je težka, saj so različni avtorji uporabljali različne podatkovne (pod)množice in različne mere uspešnosti ocenjevanja, najpogosteje povprečno napako (angl. Mean Error – ME) namesto MAE. Posledično smo se osredotočili na primerjavo z dvema standardoma, ki podajata zahteve za klinične merilnike krvnega tlaka.

Najboljši rezultati so dosegli zahteve standarda Advancement of Medical Instrumentation (AAMI). Obenem so glede na standard British Hypertension Society (BHS) za diastolični krvni tlak dosegli najvišjo oceno A in za diastolični tlak srednji oceni B in C [23].

Sistem deluje dobro za ocenjevanje diastoličnega in zmerno dobro za ocenjevanje sistoličnega krvnega tlaka. V prihodnje je potrebno dodatno testiranje predvsem s podatki, zbranimi med dnevnimi aktivnostmi, ki so zaenkrat količinsko skromni.

Chapter 1

Introduction

Blood pressure (BP) measurement is the most important commonly performed medical office test [24]. It is a direct indicator of hypertension, an important risk factor for a variety of cardiovascular diseases, which were the most common cause of death in 2015, responsible for almost 15 million deaths worldwide [1], as shown in Figure 1.1.

1.1 Motivation

Given the importance of BP, people should actively monitor it and be mindful of its changes. It is traditionally measured using an inflatable cuff, as shown in Figure 1.2. This method is still considered the "golden standard" and is preferred over digital BP monitors by doctors. It requires some effort from the user, as the sensor must be placed directly above the main artery in the upper arm area, at approximately heart height [26, 2].

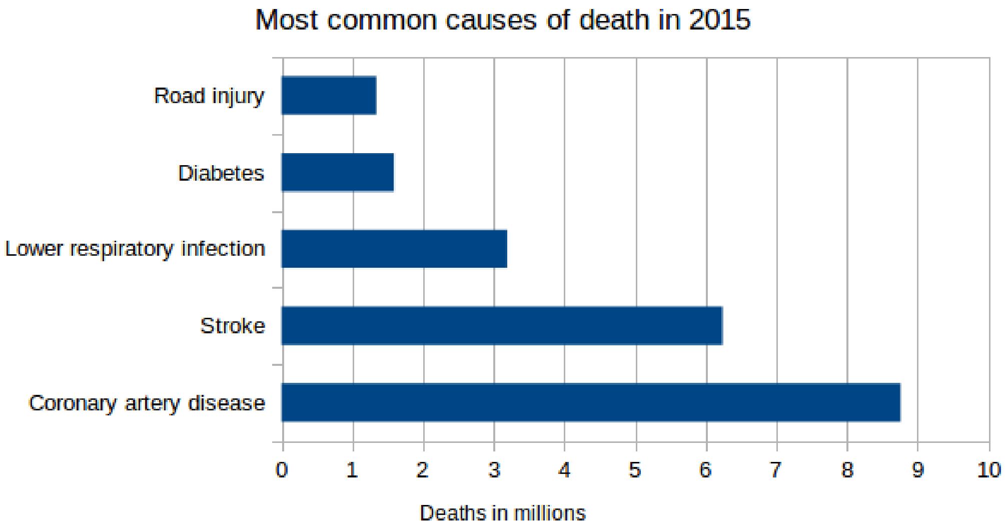


Figure 1.1: Most common causes of death in 2015, according to the World Health Organization [1]. Hypertension is the most common indicator of cardiovascular diseases such as stroke and coronary artery disease [25].

The traditional cuff-based method cannot be used during sleep and during most activities that involve movement. It may also introduce anxiety in the patient, which can affect the BP. This is commonly known as “white coat syndrome”, as this method is most often used by medical personnel [3].

Patient non-adherence has been shown [4] to be a major challenge and a barrier in ensuring effective medical treatment, sometimes even causing significant health risks as well as economic burdens. Regular BP measuring is critical when dealing with potentially hypertensive patients, as non-adherence in such cases can have fatal consequences [4]. Patients with hypertension are often given threshold BP values which must not be exceeded in order to be allowed certain activities.

Due to the combination of the factors described above, it would be useful to develop a robust mobile health (m-health) [27] system that would offer accurate and continuous BP estimation in a non-invasive way. It should ide-

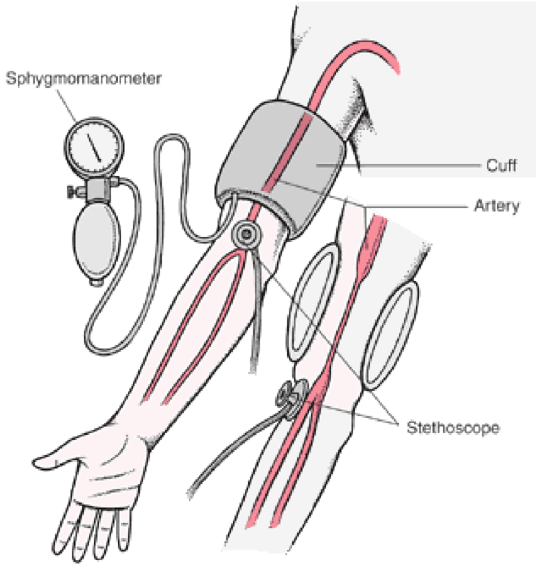


Figure 1.2: Traditional blood pressure monitoring device using an inflatable cuff [26].

ally offer accuracy in accordance with the requirements given by the British Hypertension Society (BHS) [23] and the Association for the Advancement of Medical Instrumentation (AAMI) standard [23], which require a device to have mean absolute (MAE) or mean error (ME) around 5-10 mmHg compared to a reference ground-truth BP value. Both standards are discussed in great detail in Section 5.4. Omitting the cuff would simplify the measuring process and thus likely increase user adherence as well as reduce potential anxiety. Furthermore, it would allow for measurements in cases where applying a cuff is not possible, and also increase the awareness of the user about their current medical state.

1.2 Problem description

In order to achieve the proposed low invasiveness and continuous measurement, wearable devices are considered a prime candidate. Such devices are commonly used, affordable and equipped with a plethora of sensors providing a variety of signals, however, there are also many problems related to the quality of the collected signals, as summarized at the end of Section 1.2.2.

Photoplethysmogram (PPG) is one such signal often used in modern wearables to estimate the heart rate. It requires simple and inexpensive technology and can be easily and continuously acquired in a non-invasive way by wearing a wristband. Due to the highlighted advantages, the PPG signal will be used to estimate BP in our work.

1.2.1 Photoplethysmogram (PPG)

Photoplethysmography is based on illumination of the skin and measurement of changes in its light absorption [6]. In its basic form it requires a light source (light-emitting diode – LED) to illuminate the tissue (skin), and a photodetector (photodiode) to measure the amount of light either transmitted or reflected to the photodetector. Typically, red (wavelength 645 *nm*) or green (wavelength 530 *nm*) light is used. Red penetrates the tissue deeper, while green is more robust against artefacts [5]. With each cardiac cycle the heart pumps blood towards the periphery of the body, thus producing a periodic change in the amount of light that is absorbed or reflected from the skin, as the skin changes its tone based on the amount of blood, or more precisely blood volume (BV), in the tissue [7].

As mentioned above, PPG can be measured in either transmission or reflectance mode as shown in Figure 1.3.

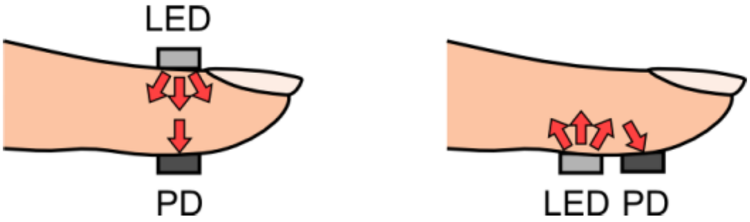


Figure 1.3: Light source (LED) and photodetector (PD) placement for both photoplethysmography modes [28].

An example of the final raw PPG signal as produced by the Empatica E4 wristband [29] is shown in Figure 1.4. It is important to note that this segment represents the PPG signal collected in a controlled environment, with the subject sitting and not moving his arm or wrist.

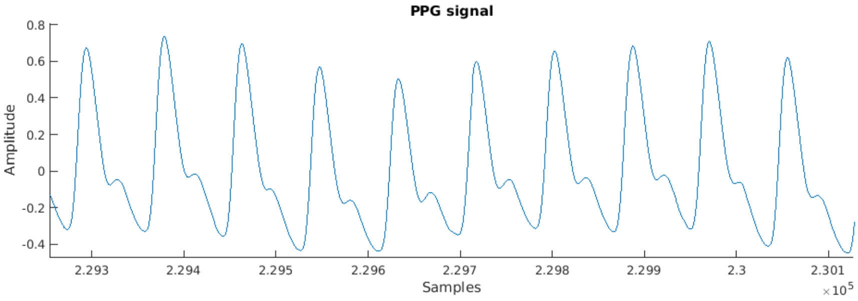


Figure 1.4: An example PPG signal as produced by the Empatica E4 wristband in a stable position, without any arm or wrist movement.

As briefly highlighted earlier, the PPG signal waveform is periodic in nature and consists of cycles, with one cycle corresponding to a single heart beat. The ideal cycle shape comprises of two peaks, systolic and diastolic. First is the systolic peak, which is higher in amplitude and corresponds to the oxygenated blood pushed from the heart traversing towards the periphery of the body. Second is the diastolic peak, which is weaker and thus lower in amplitude, and it corresponds to the deoxygenated blood returning from the peripheral point towards the heart. Two PPG cycles are shown in Figure 1.5 and the aforementioned peaks are marked.

1.2.2 BP estimation using PPG

1.2.2.1 Medical background

Xing et al. [16] discussed the medical background behind the relationship between BP and PPG. Using several assumptions, they described how blood pressure changes with blood volume due to elastic properties of blood vessels.

The elasticity of the blood vessels determines the possible amount of blood

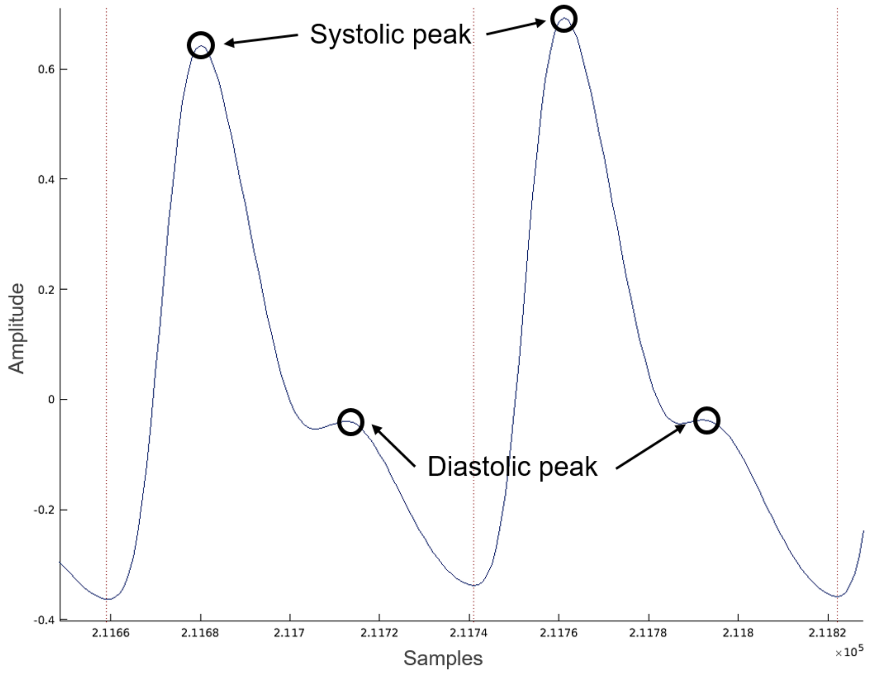


Figure 1.5: Two PPG cycles separated with vertical dashed lines. The corresponding systolic peaks and diastolic peaks are marked.

(blood volume, which is measured by PPG) and the pressure exerted on the walls of the vessels. With changes in the structure of the vessels (e.g., older people have stiffer vessels) and the response of the cardiovascular system (e.g., vessels contract or expand during stress or activity), there also come changes in both the amount of blood (shown by PPG) in the vessels, and the pressure (BP) exerted by this blood on the vessel walls. This pressure on the arterial wall is generally resisted by collagen, elastin and smooth muscle. Their combined Young's modulus of elasticity determines the stiffness of the blood vessels, which in turn influences the BP.

There are two main approaches to estimate BP from PPG. The first approach is based on pulse arrival time (PAT) or pulse transit time (PTT) and is well established [9, 10, 11]. It requires the usage of two sensors, typically an electrocardiogram (ECG) and a PPG sensor. PTT is shorter when the vessels are stiffer, which indicates a higher BP. The second approach is based on a postulated complex relationship between PPG and BP and requires the usage of PPG sensor only, and is thus less obtrusive. This approach is the subject of recent research and this thesis.

1.2.2.2 Approach using two sensors (ECG and PPG)

The first approach was proposed in 1981 by Geddes et. al. [8]. It requires one ECG sensor to be placed near the heart and another ECG or PPG sensor to be placed on a peripheral point of the body. It is based on measuring the PAT or PTT between the same R peak in both ECGs or between the R peak and the corresponding peak for the same individual pulse in the PPG signal. This time is shown to be well correlated with BP.

1.2.2.3 Approach using a single sensor (only PPG)

Recent studies from the past decade, which are discussed in detail in Chapter 2, show good correlation between PPG and BP without the use of ECG. This is very practical as it omits the requirement for an extra (commonly ECG) sensor, requiring only the collection of the PPG signal, which can be easily and continuously obtained in a non-invasive way with a wristband. Researchers who explore this approach assume the existence of a complex relation between PPG and BP, which they try to deduce in different ways. Since most studies are limited in some way, finding a general relation between PPG and BP remains an open problem which we address in our work.

By using PPG signal only, the first problem arises from the nature of the collection mechanism described in Section 1.2.1, namely the contact between the light-emitting diode, sensor and the skin. Due to this mechanism and the required high sensitivity of the sensor in order to detect tiny changes in light absorption, the signal is very prone to movement artefacts and anomalies. This issue will be resolved by extensive preprocessing with focus on artefact removal. This will be followed by precise segmentation of the PPG signal into cycles, where one cycle corresponds to one heart beat.

The unknown general relation between PPG and BP poses the second problem. To resolve it, features describing the PPG signal must be proposed and evaluated and then effectively used in a suitable regression algorithm.

1.2.2.4 Practical considerations

Finally all the considered datasets and suitable algorithms must be evaluated and the best among them implemented as a web service, which must address the problem of real time processing. The system should provide periodic updates about their BP to the user, with a relatively high frequency. This will be resolved with near-real time micro batch system architecture.

1.3 Thesis outline

The remainder of this thesis is organized as follows:

- **Chapter 2 - Related work**. Two major approaches in related work are analyzed, based on the signals used (PPG + ECG or PPG only). This thesis is placed in the context of the related work.

- **Chapter 3 - Methodology**. All modules of the system are described in detail, starting with signal preprocessing methods, continuing with feature extraction and ending with the analyzed regression algorithms.

- **Chapter 4 - Materials**. Two datasets (a clinical dataset and an everyday-life dataset) are described in detail.

- **Chapter 5 - Experiments and evaluation**. The experimental setup and evaluation procedure are discussed in detail.

- **Chapter 6 - Web service**. Implementation of the developed system as a web service, which serves the periodic BP prediction to the user's wristband.

- **Chapter 7 - Conclusions**. The work of this thesis is summarized, contributions and limitations are discussed, and future work is proposed.

Chapter 2

Related work

The following sections provide a literature overview in which we focus on the work describing previously undertaken efforts to estimate BP from PPG signal. Our main focus is on computer science and engineering methods while also analyzing some works that describe the underlying physiological mechanisms, explaining the specifics of PPG signal morphology.

In each referenced related work, we report the results in the same form as given by the authors. A common form of results, which can be found in several works discussed in the following Section, is

$$\mathrm{ME} \pm \mathrm{STD}, \tag{2.1}$$

where ME is the mean error, and STD is the standard deviation of the error. In case of using absolute errors, such as MAE [30], the notation typically omits the ± sign, and the STD is reported separately.

ME differs from MAE in the fact that the considered errors are not absolute values. This can lead to potentially negative ME, and symmetrical large positive and negative errors can cancel each other out.

In accordance with the two main approaches highlighted in Section 1.2.2, we first analyze the literature regarding BP estimation using two sensors (commonly PPG and ECG) where PTT is evaluated. We then focus on the literature regarding BP estimation from PPG signal only. The chapter is concluded by placing this thesis in the context of related work.

2.1 Approach using two sensors (ECG and PPG)

Pioneer work related to BP estimation from PPG was done in 1981 by Geddes et al. [8]. They evaluated the relationship between pulse-arrival times and diastolic blood pressure in 10 anesthetized dogs. The R peak of the ECG at heart location was used as a reference point, and the time it took for this pulse to be shown in carotid and femoral pulses was measured. They further chemically manipulated the dogs' BP and found good correlation (near linear relationship) between changes in diastolic BP and PTT.

In 1999, potential clinical applications of PTT were studied by Smith et al. [9]. They highlighted the inadequate and expensive techniques used for clinical studies of sleep disorders, which required the patient to be taken into a controlled laboratory environment. They further explained the relationship between changes in BP and sleep disorders while also reiterating the good correlation between PTT and BP. PTT was praised due to its simplicity and low cost, and its potential uses were proposed.

Poon et al. [10] proposed explicit equations for systolic (SBP) and diastolic BP (DBP) based on Moens–Korteweg formula, which was derived in the 19th century and models the relationship between pulse wave velocity (PWV) and the incremental elastic modulus of the arterial wall. They conducted experiments with 85 subjects and obtained promising results with ME as low as 0.6±9.8 mmHg. The evaluation method using ME instead of MAE is questionable, since it may display superior results compared to actual performance of the model, as large symmetrical positive and negative errors can cancel each other out in the final reported ME.

Recent work was published by Kachuee et al. [11] in 2017 in which BP was estimated using PTT from ECG and PPG signal. They evaluated several regression methods, such as Linear Regression, Random Forest, Support Vector Machine (SVM), etc. They achieved MAE of 11.17 mmHg for SBP and 5.35 for DBP, with the corresponding STD of 10.09 mmHg for SBP and

6.14 mmHg for DBP, using AdaBoost regression. They reported their results meet the Association for the Advancement of Medical Instrumentation (AAMI) and the British Hypertension Society (BHS) standards for DBP, however, SBP has proven to be more difficult to estimate.

Further research has been conducted using the PTT approach, showing promising results. In recent years, however, the focus of researchers is being shifted towards BP estimation using only the PPG signal, due to the requirement of two sensors to measure the PTT being more obtrusive.

2.2 Approach using a single sensor (only PPG)

As mentioned previously, a great deal of recent research effort has been directed towards BP estimation using only the PPG signal. Obtaining the PPG signal typically requires the user to wear a small simple device with an LED light and a photodetector. In a hospital or laboratory setting, such a device is commonly placed on the tip of the finger or in the earlobe area, as the tissue there is rather translucent. Despite the requirement for only a single sensor, wearing such a device can still be obtrusive, due to its location. It would be highly beneficial, if the PPG signal could be obtained in a less obtrusive way, allowing the user to conduct most activities without limitations.

It has recently become common for the PPG collection sensor to already be embedded in most popular wristbands (e.g. Apple Watch [31], Microsoft Band 2 [32], etc.), as it is so simple and inexpensive to implement. Furthermore, this approach is becoming increasingly popular with the recent surge of m-health wearable devices and applications, as users are very comfortable with using such devices [27]. A wristband seems like an optimal wearable device for PPG collection, since it does not limit the user in almost any activity. Despite this, research dealing with BP estimation using a wristband is scarce, while research dealing with BP estimation using only the PPG signal is more plentiful.

An early attempt at PPG only approach was conducted by Teng et al. [12] in 2003. They examined the relationship between arterial blood pressure (ABP) and certain features of the PPG signals obtained from 15 young healthy subjects. The data was collected in a highly controlled laboratory environment, ensuring constant temperature, no movement and silence. Using correlation analysis four best features were chosen and used in a linear regression model to predict the BP. The ME between the estimated and the measured blood pressure were 0.21 ± 7.32 mmHg for SBP and 0.02 ± 4.39 mmHg for DBP. Again, these results are given with ME instead of MAE, which might not reflect the actual performance of the derived model.

A paper was published in 2013 by Lamonaca et. al [13] in which they used data from Multiparameter Intelligent Monitoring in Intensive Care (MIMIC) waveform database [22] to extract 21 time domain features and use them as an input vector for artificial neural networks (ANN). The data was obtained from a higher number and variety of patients in a less controlled environment compared to previous work. Patients from the MIMIC database were treated in a hospital environment where the signals were measured during their stay without any strict movement, temperature or sound restrictions. The PPG was measured with a fingertip hospital device, while ABP was measured invasively using a catheter in the artery. The authors defined some additional time domain features that were not commonly used before and are shown to describe the PPG cycle shape well. They reported lowest MAE of 3.80 mmHg for SBP and 2.21 mmHg for DBP, with corresponding STD of 3.46 mmHg for SBP and 2.09 mmHg for DBP, meeting the AAMI standards. The exact data used from the MIMIC database is not disclosed and the research is still based solely on data collected in a hospital, using hospital equipment, albeit with less strict limitations regarding the patients.

As user and manufacturer attention shifts increasingly towards mobile devices, several approaches were proposed which rely on such devices for BP estimation.

Lamonaca et al. [14] published another research in 2013 in which they used a smartphone camera to capture the PPG signal using the camera flash as the light source and phone camera as the photodiode. PPG features were again extracted and fed to a neural network which estimated the SBP and DBP. All the data processing and BP evaluation was done in a cloud in order to reduce the computational burden on the device. The results were again promising with the maximum error not exceeding 12 mmHg, however, such a method requires additional user effort, as the user must place and hold his finger over the camera and LED light. This prevents any other activities during this time.

In a recent 2015 paper, Banerjee et al. [15] emphasized the importance of signal preprocessing in order to greatly improve the BP estimation accuracy. They suggested several signal improvement methods such as filtering, baseline drift removal and cycle selection based on cycle quality assessment. They further proposed additional features based on Gaussian modelling of each cycle. These features were then used in a neural network and improvements were shown in comparison to their previous work on the same data, which consisted of signals gathered by 15 subjects using their phones. Similar limitations apply, as the users had to hold their finger over the camera and LED light of the phone.

Xing et al. [16] were the first to propose a normalization algorithm, for which they claim that it removes subject-specific or device-specific contributions to the PPG signal. This makes the BP estimation algorithm completely independent of subject or device. Amplitudes and phases are extracted from the FFT transformation of the PPG signal and used to train an ANN. Their approach was validated using 69 patients from the MIMIC II database. They reported good correlations between the predicted and actual BP values, where 86% of SBP and 95% of DBP cases had absolute error $\leq$ 10 mmHg.

2.3 Related work summary

Tables 2.2 and 2.3 presents a summary of the related work described in the previous sections. The related work is grouped based on the two BP estimation approaches mentioned earlier, and the following key perspectives are identified and summarized where applicable:

- The data that was used for the evaluation. How much and what type of data was used, was it collected from a constrained or an unconstrained environment.
- The methods that were used to derive the relationship model between input signals and the BP. Which method was chosen based on the best results.
- The reported results. Which metric was used for the evaluation, what were the best results.

Approach using two sensors (ECG and PPG)			
Study	**Data**	**Methods**	**Reported results**
Geddes et al. [8]	ECG signal alongside reference BP from dogs in constrained env.	Correlation analysis between changes in PTT and BP	Good, near linear correlation of changes in PTT and BP
Smith et al. [9]	ECG and PPG signal alongside reference BP from unknown nr. of subjects in constrained env.	Analysis of potential applications of PTT based on established research	PTT proposed as a promising alternative to current complicated tests
Poon et al. [10]	ECG and PPG signal alongside reference BP from 85 subjects in constrained env. with calibration	Explicit equations for SBP/DBP derived from Moens-Korteweg formula	ME of 0.6±9.8 mmHg for SBP and 0.9±5.6 mmHg for DBP
Kachuee et al. [11]	ECG and PPG signal alongside reference BP from MIMIC II database from roughly 1000 unique subjects in a constrained env.	PTT and additional features used in 5 regression algorithms	MAE 11.17 for SBP and 5.35 for DBP with STD of 10.09 mmHg for SBP and 6.14 mmHg for DBP, using AdaBoost

Table 2.2: Summary of related work based on the two sensors approach.

Approach using a single sensor (only PPG)			
Study	**Data**	**Methods**	**Reported results**
Teng et al. [12]	PPG signal alongside reference BP from 15 subjects in highly constrained env.	Best of 4 features used in linear regression	ME of 0.21±7.32 mmHg for SBP and 0.02±4.39 mmHg for DBP
Lamonaca et al. [13]	PPG signal alongside reference BP from MIMIC database in constrained env.	Extracted 21 time domain features and fed them in an ANN	ME of 3.80±3.46 mmHg for SBP and 2.21±2.09 mmHg for DBP
Lamonaca et al. [14]	PPG signal from phone camera and reference BP from unknown nr. of subjects	Extracted 21 time domain features and fed them in an ANN	Maximum error $<$ 12 mmHg

Banerjee et al. [15]	PPG signal from phone camera and reference BP from 15 subjects in constrained env.	PPG signal modelled with a sum of 2 Gaussian functions, denoising, and temporal features computed and fed into an ANN	Notable improvements compared to their previous work.
Xing et al. [16]	PPG signal and reference BP from 69 patients from MIMIC II database	Normalization of PPG, amplitudes and phases extracted from the FFT of the PPG and then fed into an ANN	86% of SBP and 95% of DBP with absolute error $\leq$ 10 mmHg

Table 2.3: Summary of related work based on the PPG only approach.

2.4 Thesis in the context of related work

There are several common themes to the related work we have reviewed.

Firstly, the PTT approach is well established and understood, however, due to the requirement of two sensors, recent research effort shifts towards the single sensor PPG only approach, which is less obtrusive and more suitable for modern wearable devices.

Secondly, the data used in experiments is often limited to a low number of subjects or limited by the strict collection process requirements imposed upon the subjects in constrained laboratory or hospital environments. To our current knowledge, an everyday-life dataset collected with a wristband has never before been thoroughly analyzed.

Finally, substantial progress has been made in BP estimation using mobile smartphone cameras with LED flash, however, this limits the user and requires his full attention for the entirety of the estimation process and is not suitable for use during many activities.

This thesis aims at developing a general approach which can be used on any input data. Accordingly, two datasets will be used and evaluated:

1. All the patients having both PPG and BP signal from the MIMIC database are considered [22] as the clinical dataset.

2. A custom dataset is being collected by as many subjects as possible during their everyday activites, using the Empatica E4 wristband.

It is important to note that a replication of the results reported by some of the current state of the art approaches reviewed in Chapter 2 was attempted with all eligible MIMIC database data. The results as given in some related work could not be replicated. We suspect that in some cases, a subset of all the available MIMIC data was chosen in the related work. We tried contacting some of the authors regarding their work, but we had limited success in obtaining the details regarding their work. In other cases, the results might be replicated, however, the reported error metrics are considered inadequate.

We consider reporting only the ME as inadaquate, since it might not reflect the actual performance of the derived model very well, as symmetrical positive and negative errors cancel each other out.

This indicates that a general approach to BP estimation using only PPG signal is still required and will be proposed in this thesis.

Chapter 3

Methodology

The proposed system for BP estimation consists of two main modules, namely the signal-preprocessing and the machine-learning module. The former module is responsible for producing segmented high-quality PPG signal cycles as the output, given an arbitrary noisy PPG signal as the input. Each PPG cycle on the output corresponds to a single heart beat, and should have minimal or no morphological alternations due to artefacts. The latter module first extracts a number of features describing the PPG signal on a per-cycle-basis. A subset of relevant features is then determined using the RReliefF algorithm [19]. Finally, the performance of several regression algorithms is evaluated using several experiments. The best among them is chosen for the creation of predictive models for SBP and DBP. These two modules are complemented by the web service, which allows for the interaction between the user's wristband and the predictive model. The proposed system architecture is summarized in Figure 3.1.

In this chapter, we overview the methods used in our work. We focus on the ideas behind each method, while the exact parameters, experiments and results, are discussed in Chapter 5.

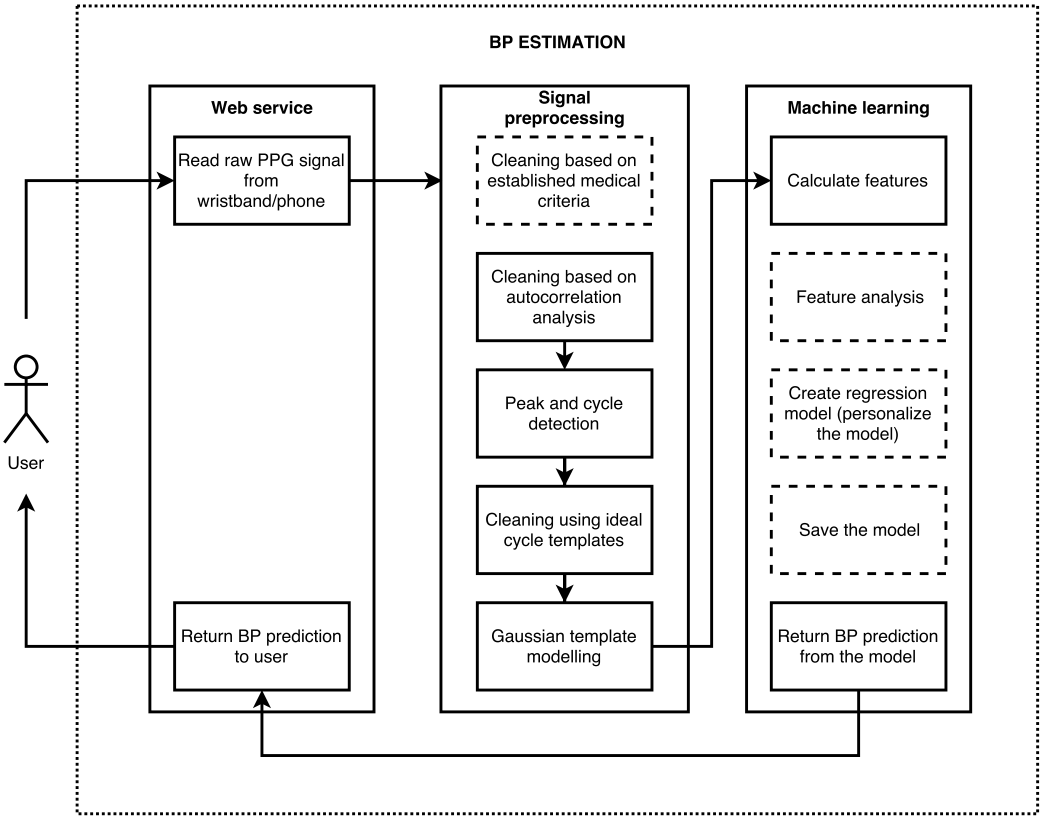

BP ESTIMATION
Web service
Read raw PPG signal from wristband/phone
Return BP prediction to user
Signal preprocessing
Cleaning based on established medical criteria
Cleaning based on autocorrelation analysis
Peak and cycle detection
Cleaning using ideal cycle templates
Gaussian template modelling
Machine learning
Calculate features
Feature analysis
Create regression model (personalize the model)
Save the model
Return BP prediction from the model
User

Figure 3.1: Proposed architecture of the BP estimation system. The rectangles with the dashed lines correspond to steps, that are only executed once or periodically (initial model creation or personalization).

3.1 Signal preprocessing

As mentioned earlier, the PPG sensor must be highly sensitive in order to detect tiny variations in light absorption of the measured tissue. This in turn makes the sensor highly susceptible to movement artefacts, which can notably distort the signal. This problem is exacerbated by using a wristband as the measuring device, since the contact between the sensor and the skin can be compromised during arm movements. The use of green light partially alleviates the problem, as described in Section 1.2.1, however, notable artefacts often remain in the signal. Subsequently, substantial effort is directed towards the PPG preprocessing in order to obtain the signal of the highest possible quality.

3.1.1 Cleaning based on established medical criteria

The possible amplitude values of the PPG signal are not medically determined, but rather dependent on the specific device [33, 34], while the ABP values are known to only be possible within a certain range (e.g., between 0 mmHg and 300 mmHg, etc.). We can thus typically eliminate some segments in each ABP recording due to the signal values being extreme and beyond anything realistic. This is often the result of unexpected movement of the catheter in the artery, or removal/replacement of the catheter. It is thus advisable to first inspect the ABP signal and remove these segments. Extreme values of BP (e.g., over 300 mmHg) are clinically established to be impossible, and segments with such values can be safely removed.

A 5-second sliding window with no overlapping is used to detect extreme BP values or extreme changes of BP in a short time period. Thresholds for extreme values and changes are selected based on established medical criteria as given in related work [16]. Some criteria were slightly modified in accordance with empirical observations in our data, and are summarized in Table 3.1. We have made the criteria slightly less strict, as large parts of our data would be excluded by the original criteria (e.g., the original criteria excludes all data with SBP > 180, while we observed some segments with SBP between 180 mmHg and 200 mmHg).

3.1.2 Cleaning based on autocorrelation analysis

As mentioned before, both the PPG and ABP signal are periodic in their nature. Using autocorrelation analysis, it is expected for the maximum autocorrelation of a signal segment (a short excerpt of the signal) to be rather high, if the signal is actually periodic in this segment. A low maximum autocorrelation indicates a lack of a periodic pattern in this segment, thus

Criterion	**Threshold**
SBP	> 220 or < 60
DBP	> 150 or < 30
SBP – DBP	< 20
ΔSBP or ΔDBP	> 50

Table 3.1: Modified established medical criteria and thresholds for rough signal cleaning. The ΔSBP and ΔDBP signify a change of BP value between two subsequent 5-second PPG segments. All 5-second segments meeting any of these criteria are removed from the signal.

indicating an exceptionally noisy segment.

Autocorrelation analysis measures the correlation of a signal with a delayed copy of itself. It is defined as the correlation between y_i and y_{i+k}, where $k = 0, ..., K$. K is the maximum lag in samples at which the sample autocorrelation function (SACF) is computed. SACF at lag k is defined as:

$$r_k = \frac{c_k}{c_0}, \tag{3.1}$$

where

$$c_k = \frac{1}{N}\sum_{i=1}^{N-k}(y_i - \bar{y})(y_{i+k} - \bar{y}), \tag{3.2}$$

c_0 is the sample mean given as:

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i, \tag{3.3}$$

and N is the total number of samples in the given time series [35].

Again, a 5-second sliding window was used and the SACF within the window was computed. It is expected that a 5-second window contains from 3 (during sleep) to 9 (during intense activity) heart beats [36], whose morphology should be similar, and the maximum autocorrelation should be relatively high. A threshold was determined empirically. It was set to be notably lower than the average maximum autocorrelation of the segments containing clean periodic signal, as only the most noisy signal should be removed in this phase.

An example is shown in Figure 3.2, where the top-left subplot shows an example 5-second PPG segment, and the bottom-left shows this same segment with added white Gaussian noise. The right subplots show the corresponding autocorrelations at maximum 1-second lag (125 samples with 125Hz sampling frequency (F_s)).

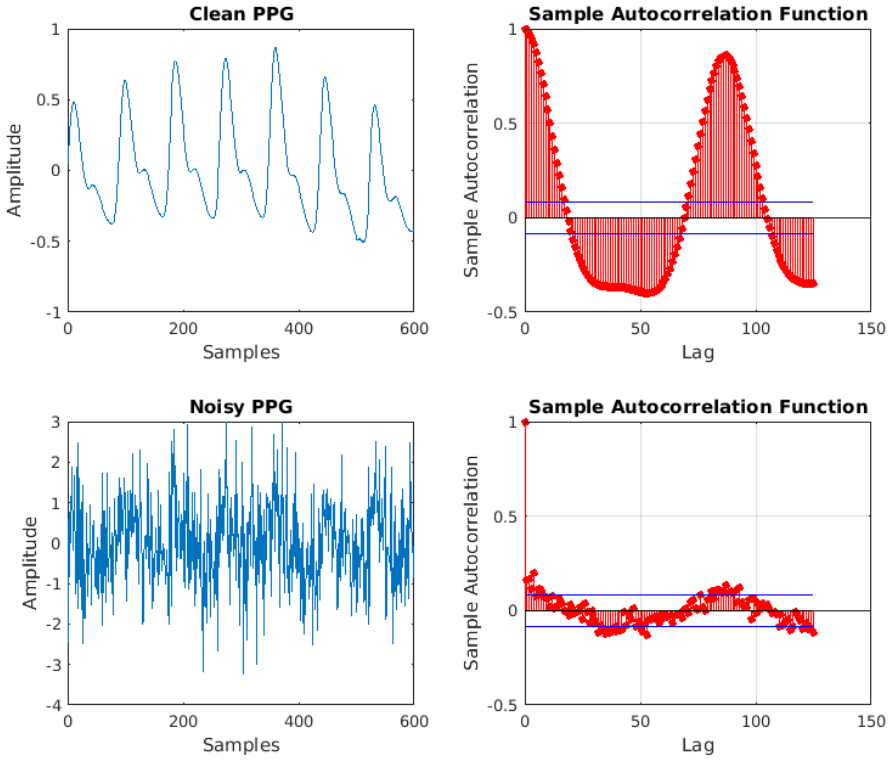


Figure 3.2: Top-left subplot shows an example PPG segment obtained from a real patient in a hospital environment, and the top-right shows the obvious peaks in autocorrelation, corresponding to one cycle lag. The bottom-left subplot shows the same segment with added white Gaussian noise, while the bottom-right shows the decreased autocorrelation of such a noisy segment.

It can be clearly seen that the autocorrelation maximum for the clean signal segment reaches a peak at just under 100-sample lag, which corresponds to one cycle. The autocorrelation maximum value is close to 1, as the signal is highly periodic. On the other hand, the maximum autocorrelation value for a noisy signal segment is extremely low, at just over 0.1. The average maximum autocorrelation value of a realistic noisy PPG segment obtained from a wristband is around 0.5. We have thus empirically determined a threshold of 0.8 which must be met in order to not discard a PPG segment.

It is important to note that the autocorrelation starts at value 1, corresponding to 0-sample lag and complete autocorrelation, meaning identical signal segment. With very small values of lag, the autocorrelation typically remains high, even for a slightly noisy signal. Thus, we search for the maximum value only after 13-sample lag, which corresponds to 0.104-second lag (at $F_s = 125$Hz).

3.1.3 Peak and cycle detection

In order to do continuous BP estimation, the BP could be estimated on a per-cycle basis. Subsequently, we should derive features that describe each individual PPG cycle, as is common in the vast majority of the related work discussed in Chapter 2. One PPG cycle corresponds to a single heart beat. A robust segmentation of the PPG signal into cycles is thus mandatory.

We have used a two-phase pulse detector algorithm proposed by Lazaro et al. [17] to detect the prominent PPG systolic peaks. The first phase is a linear filtering transformation and the second phase is a dynamic thresholding operation. After the peaks are detected, finding the cycle start-end locations, which correspond to the dominant valleys, is simpler.

3.1.3.1 Linear filtering transformation

The purpose of the filtering transformation is to enhance the prominence of the abrupt upslopes of systolic PPG peaks over the smoother upslopes of diastolic or dicrotic peaks. This ensures that only systolic peaks are considered, while diastolic peaks are ignored.

The transformation is based on a linear-phase finite impulse response (FIR) low-pass-differentiator (LPD) filter.

The key characteristic of a FIR filter is that its response to any finite input is of finite duration. The response settles to zero in a finite time, as it has no internal feedback. Such a filter is also said to be non-recursive. FIR filters are commonly linear-phase, meaning that they shift all the frequency components of the input signal in time by the same constant amount. This constant amount is known as the phase delay. Importantly, this does not cause phase distortion, meaning the shape of the waveform is preserved [37].

The idea behind low-pass filter is to remove all the high frequency components from the signal and allow only the low frequencies in the specified range to pass.

A differentiator is a filter designed such that the output of the filter is approximately directly proportional to the rate of change (the time derivative) of the input. The rate of change corresponds to the slope of the waveform and is the highest at the systolic rise area of each PPG cycle.

3.1.3.2 Dynamic thresholding

In order to avoid the detection of potential double peaks, and more importantly the detection of diastolic peaks, dynamic thresholding is used. Let us denote the threshold value at point n as $y_t(n)$, and the peaks in the filtered signal as n_A. A time-varying threshold is used between the detections of n_A. The threshold keeps the value of the previous detected peak $n_{A_{i-1}}$ during a refactory period lasting for 150 ms or $N_r = 0.15{\cdot}Fs$ samples. After this period, it begins to gradually decrease in a linear manner. If no new peak detection occurs in a time period $\hat{m}_{AA_i}$, then the threshold will drop to a percentage $\alpha < 1$ of the previous detected peak $n_{A_{i-1}}$, and then maintain its value. The described logic can be formulated as

$$y(n) = \begin{cases} y(n_{A_{i-1}}) & ; (n - n_{A_{i-1}}) < N_r \\ \frac{(\alpha-1)y(n_{A_{i-1}})}{\hat{m}_{AA_i} - N_r}(n - n_{A_{i-1}} - N_r) + y(n_{A_{i-1}}) & ; N_r \leq (n - n_{A_{i-1}}) < \hat{m}_{AA_i} \\ \alpha y(n_{A_{i-1}}) & ; (n - n_{A_{i-1}}) \geq \hat{m}_{AA_i} \end{cases}, \tag{3.4}$$

where

$$\hat{m}_{AA_i} = \text{median}\{(n_{A_{i-4}} - n_{A_{i-3}}), (n_{A_{i-3}} - n_{A_{i-2}}), (n_{A_{i-2}} - n_{A_{i-1}})\}. \tag{3.5}$$

Next, the detected peaks n_A from the FIR-LPD filtered signal are used in the original PPG signal to determine the location of the steepest abrupt upslope, corresponding to the PPG systolic rise. Then the maximum peak is detected in the 300-ms interval in the PPG starting at the location of n_A. The actual systolic PPG peaks are found and marked as n_A^*. This final procedure is shown in Figure 3.3.

Finally, we determine the indices of the cycle start-end locations, which correspond to the dominant valleys in the interbeat interval (IBI) between the peaks. A simple algorithm is employed, which moves sample by sample between the detected peaks. In this area of the signal, it searches for the last valley before the next systolic peak. The search area is limited to 70% of the whole IBI by ignoring the first 20% and last 10% of the IBI, as shown in Figure 3.4. These values were chosen empirically, as it never happens in our data, that a true cycle start-end valley is located just a few samples (10% of IBI before or 20% of IBI after) from the systolic peak, while an anomaly valley could be located there. A larger amount (20%) is ignored at the start of the IBI compared to the end (10%), since the morphology of a PPG cycle is such that the part of the cycle after systolic peak is always longer compared to the part before the systolic peak, as seen in Figure 3.4.

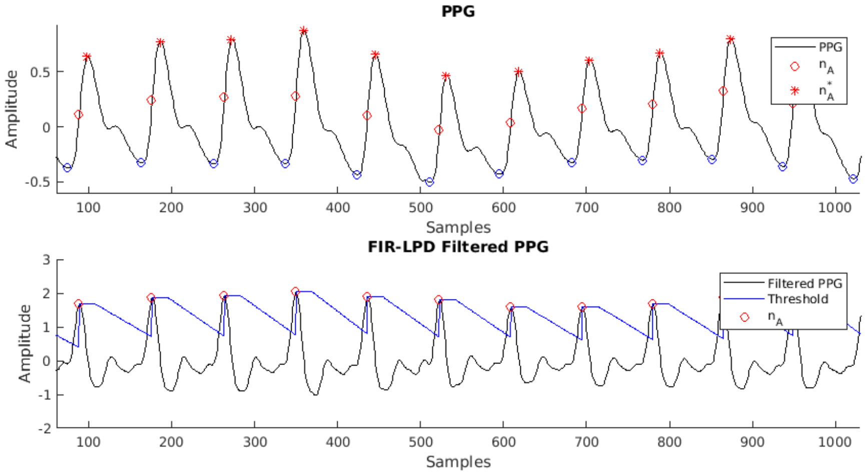


Figure 3.3: Systolic peak detection procedure. The top subplot shows a PPG segment, while the bottom subplot shows the FIR-LPD-filtered version of the same segment. The peaks in the FIR-LPD-filtered PPG, denoted as n_A, correspond to the steepest systolic rise areas in the PPG.

3.1.4 PPG cleaning using cycle templates

As mentioned earlier, artefacts and noise in the PPG signal are a major problem. In order to be able to use features that describe the signal on a per-cycle basis, only high-quality cycles should be considered.

We propose a cleaning procedure based on the work of Li et al. [18] that creates an average cycle template T from all the cycles in a 30-second window. The window was chosen to be long enough to contain a notable amount of cycles for template creation, while being short enough to allow for continuous estimation of BP. A prediction every 30 seconds is nearly continuous, as BP does not typically change very abruptly, except in extreme cases such as arterial bleeding. The correlation of each individual cycle with the template is then computed, and which cycles should be kept or discarded is determined. This is based on two assumptions. First, we expect a useful average 30-second segment of the PPG signal to contain a notable number of cycles without artefacts or noise, otherwise the segment is too noisy to be of use. Second,

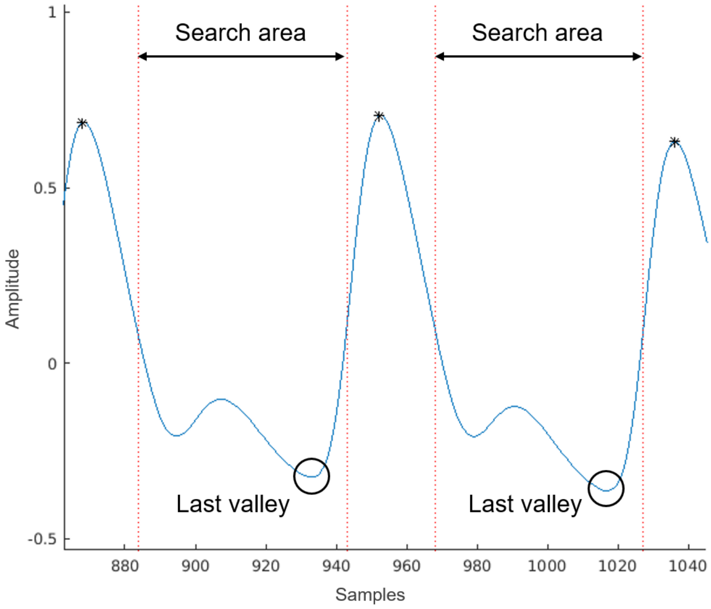


Figure 3.4: Cycle start-end locations detection, which correspond to the dominant valleys in the IBI.

we assume that the very noisy segments of the PPG signal do not contain prolonged periodic wave-like patterns.

3.1.4.1 Average cycle length

In order to create the template by averaging all the cycles in a 30-second segment, the same number of samples for each cycle must be taken. In reality, cycle length can vary by a few samples from cycle to cycle. At average PPG sensor sampling frequencies ranging from 16Hz (lower quality sensor) to 64Hz (higher quality sensor), this results in a few hundredths of a second difference in duration. We resolve this discrepancy in cycle lengths by determining the most likely cycle length in the current segment. This is once again achieved using the autocorrelation analysis described in Section 3.1.2.

Considering the periodic nature of the PPG signal, we can expect the first autocorrelation peak to be located at one-cycle duration lag. This can be clearly seen in Figure 3.2. This number of samples L is taken as the most likely length of a cycle for a given 30-second segment.

3.1.4.2 Template generation

Once the cycle start-end locations are known, L samples are taken from each cycle starting point. Considering that L is the expected length of a cycle for a given segment, all the key morphological characteristics, most notably the systolic and diastolic peak, are preserved for each cycle. The template is then generated by taking the mean of cycle values at the same time.

An example template resulting from the described procedure is shown in Figure 3.5.

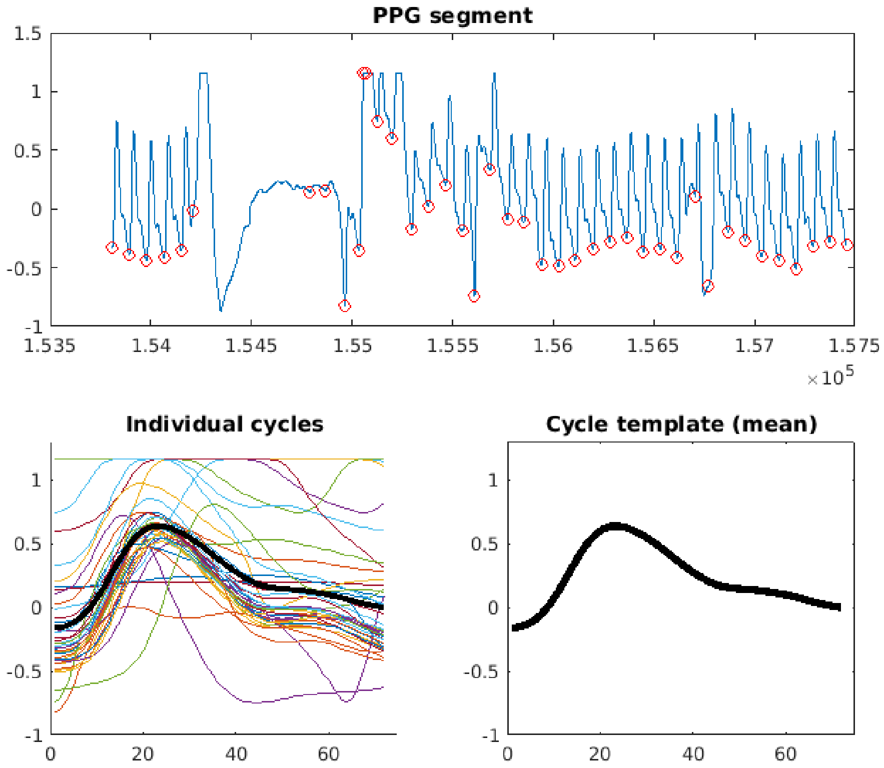


Figure 3.5: The top subplot shows a random PPG segment. The bottom-left subplot shows the individual detected cycles in this segment, and the bottom-right shows the computed template of all the cycles, which is computed as the mean of cycle values at the same time.

3.1.4.3 Cycle quality assessment

Once an initial cycle template is created for a given segment, each individual cycle within this segment is compared with the template. Each cycle quality is assessed using several signal quality indices (SQIs), which are defined as follows:

1. SQI_1: direct linear correlation using the Pearson's correlation coeffi-

cient, given as

$$\rho(A,T) = \frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{\overline{A_i - \mu_A}}{\sigma_A}\right)\left(\frac{T_i - \mu_T}{\sigma_T}\right), \tag{3.6}$$

where A and T are a cycle and the template, μ_A and σ_A are the mean and standard deviation of A, μ_T and σ_T are the mean and standard deviation of the template, and N is the total number of samples [38]. The length of each cycle is fixed to L samples from it's starting point in order to be able to compute the correlation to the template of the same length.

2. SQI_2: direct linear correlation, again using the Pearson's correlation coefficient, as defined in Equation 3.6, however, this time each cycle is linearly resampled to length L, using piecewise linear interpolation.

3. SQI_3: The correlation between each time-warped cycle and the template, as given by Dynamic Time Warping (DTW) [39], using the Euclidian distance metric. DTW allows for non-linear time series matching, meaning it will recognize two signals having the same characteristics delayed in time, as shown in Figure 3.6.

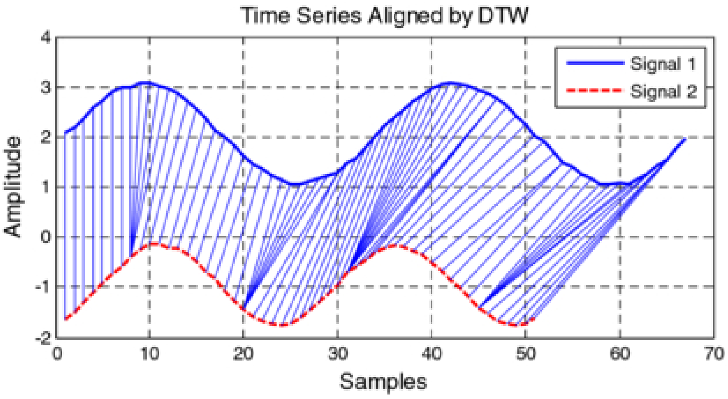


Figure 3.6: Dynamic time warping example for two time series.

If we consider two cycles A and B of lengths n and m as

$$\begin{aligned} A = a_1, a_2, ..., a_n \text{ and} \\ B = b_1, b_2, ..., b_m, \end{aligned} \tag{3.7}$$

then to align these sequences of different lengths using DTW, a matrix D is first constructed. The element of the matrix at location (i,j) contains the distance $d(a_i, b_j)$ between the points a_i and b_j. The aim of the DTW is to find an optimal path from (0,0) to (n,m), which minimizes the sum of the distances on the path. This total distance is inversely proportional to the similarity between the two signals, and the warping path through the matrix of distances tells us the optimal time warping of the signals.

An example of this procedure using a PPG cycle and a template is shown in Figures 3.7 and 3.8.

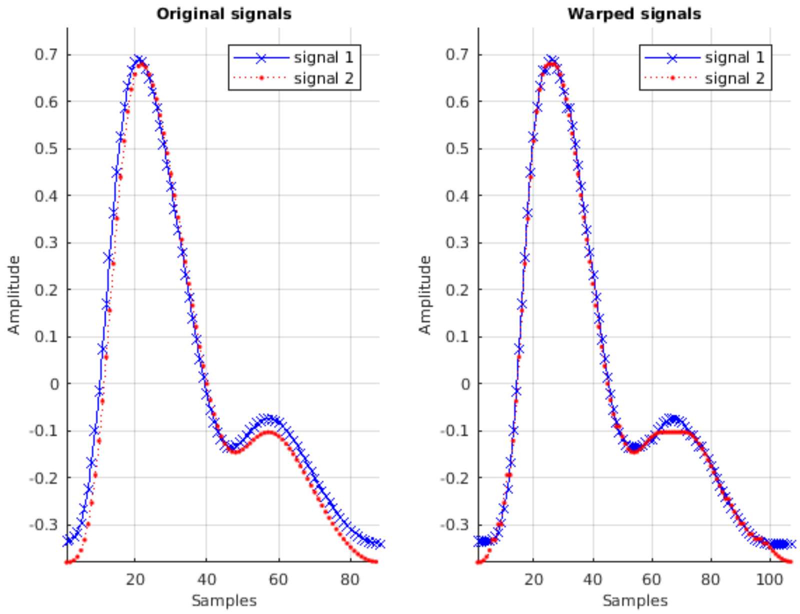


Figure 3.7: The left subplot shows the original PPG cycle and template. The right subplot shows the time-warped versions. We can see in the right subplot that signal 2 is warped in the area around the diastolic peak in order to increase the fit of the downslope.

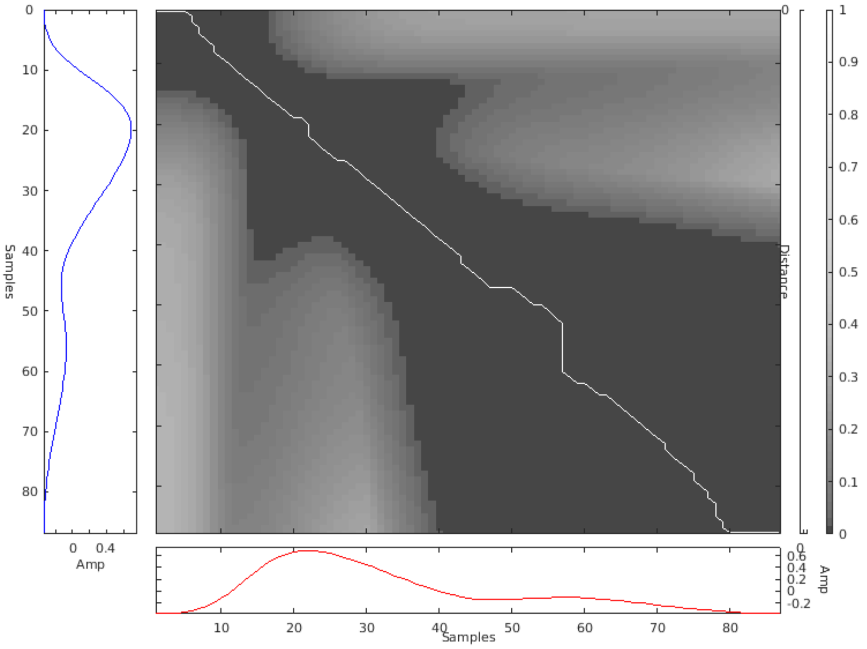


Figure 3.8: The matrix showing a warping path. The rows correspond to samples of the blue signal on the left, while the columns correspond to samples of the red signal at the bottom. Darker squares signify smaller distance, while lighter squares signify greater distance. The white line is the warping path. An obvious warp denoted by the vertical line in the warping path is seen between 50th and 60th sample of the blue signal. It signifies that these samples of the blue signal are all compared to the signle sample of the red signal. This is also seen in Figure 3.7.

Thresholds for all three SQIs are set in accordance with related work [18] and those cycles that meet the required thresholds are kept, while others are discarded. Additionally, if more than half of detected cycles in a 30-second PPG segment are discarded, then the whole segment is considered unreliable and is discarded in entirety. The successful removal of artefacts as the result of the described procedure is shown in Figure 3.9.

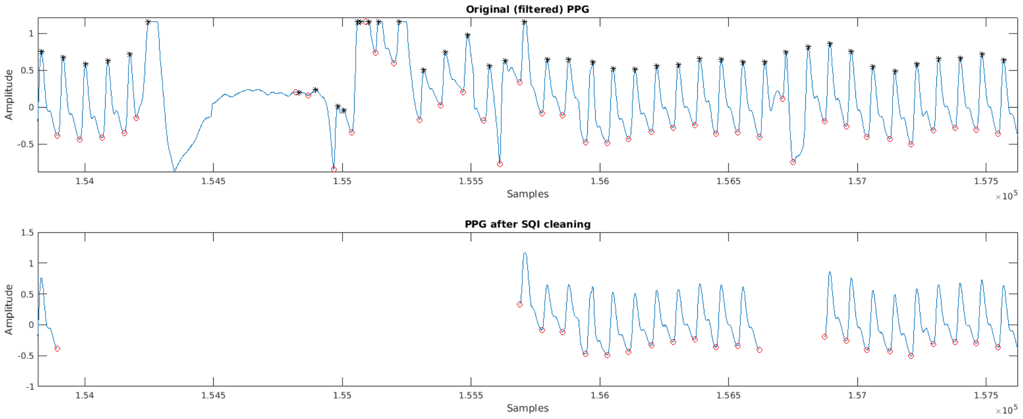


Figure 3.9: The top subplot shows a random PPG segment with an obvious artefact. The bottom plot shows this same segment with the artefact successfully removed by our cleaning procedure.

3.1.5 Gaussian template modelling

The ideal shape of a PPG cycle waveform, as detailed in Section 1.2.1 and shown in Figure 1.5, is expected to contain two peaks. The higher-amplitude systolic peak should be followed by a lower-amplitude diastolic peak. More than two peaks present in a cycle signify either an anomaly in the signal or an incorrectly detected cycle. On the other hand, it is common for a cycle in our data to contain just a single true peak, corresponding to the systolic peak, while the diastolic peak is often not present or cannot be clearly seen in the waveform. As the diastolic peak is important for the description of the waveform, we attempted to obtain only the cycles of expected shape, having the diastolic peak expressed.

It was shown by Banerjee et al. [15] that a PPG cycle can be suitably modelled for the purpose of BP estimation with a sum of two Gaussian functions given as

$$y_G(n) = a_1 * e^{-(\frac{n-b_1}{c_1})^2} + a_2 * e^{-(\frac{n-b_2}{c_2})^2}, \tag{3.8}$$

where $y_G(n)$ is the Gaussian modelled cycle value, n is a given cycle sample, a_1 and a_2 are the first and second peak amplitudes, b_1 and b_2 are the first and second peak locations, and c_1 and c_2 are the first and second peak widths.

Using empirical analysis of our data alongside the expert knowledge about the PPG cycle morphology obtained from the related work, we have determined restrictions for each of the six parameters that must be met. We have set the restrictions as follows:

1. the systolic peak must precede the diastolic peak ($b_1 < b_2$),
2. the systolic peak amplitude must be larger than diastolic peak amplitude ($a_1 > a_2$),
3. the width of both peaks must be suitably small for them to be clearly shown.

All the parameters with the corresponding explicit restrictions are given in Table 3.2.

First Gaussian	
Parameter	**Restriction (From - To)**
a_1	$0.8 * \max(T) - 1.2 * \max(T)$
b_1	$0.1 * L - 0.4 * L$
c_1	$0.05 * L - 0.15 * L$
Second Gaussian	
Parameter	**Restriction (From - To)**
a_2	$0.2 * \max(T) - 0.6 * \max(T)$
b_2	$0.5 * L - 0.9 * L$
c_2	$0.05 * L - 0.15 * L$

Table 3.2: The explicit restrictions for each of the six parameters defining the Gaussian model of a PPG cycle waveform 3.8. T is the current template, $max(T)$ is the maximum amplitude, and L is the length of the current template.

Once the parameter restrictions were determined, each cycle template was modelled with the sum of two Gaussians with the restrictions applied. The coefficient of determination R^2 was used to measure the goodness of the fit. It tells us the proportion of the total variation that is explained by the model and is defined as

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}, \tag{3.9}$$

where SSR is the sum of squares regression, which is the sum of the squared differences between the prediction for each observation and the mean of all samples. SSE is the sum of squares of error, meaning the sum of all squared prediction errors. SST is the sum of squares total, meaning $SSR + SSE$.

R^2 can take values between 0 and 1. If a given cycle was close to the expected shape, having visible systolic and diastolic peak in the expected order, the R^2 is high and close to 1. Otherwise, the R^2 is low. A threshold was empirically set at 0.9, which dictates a high matching requirement between the cycle template and the restricted Gaussian model. An example is shown in Figure 3.10.

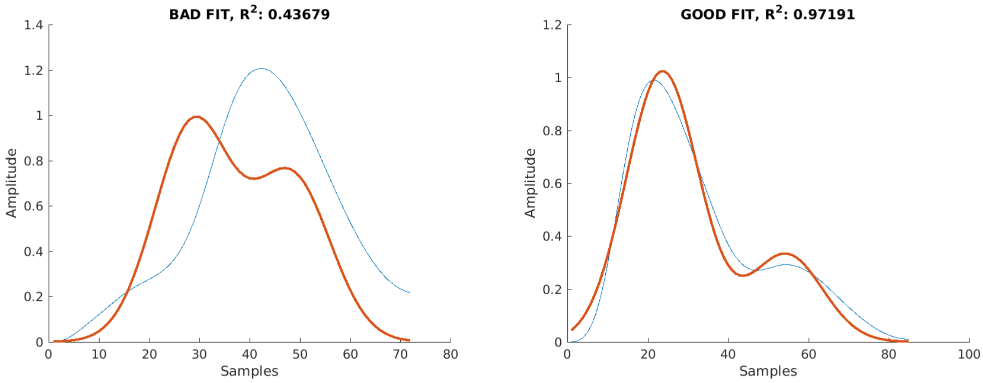


Figure 3.10: Gaussian modelling of PPG templates with restrictions as given in Table 3.2.

3.2 Feature analysis

In machine learning, a feature is a measurable property or characteristic of the observed phenomenon [40]. In our case, the features should describe the PPG waveform and should include the underlying connection between the PPG and BP.

3.2.1 PPG amplitude

The term ”plethysmogram” is derived from the Greek root ”plethysmos”, meaning ”to increase”. The signal is not given a unit designation, and the amplitude cannot be used to compare one patient waveform to another, as it is typically device-specific and related to the auto-gain found in most PPG measuring devices [33]. Subsequently, a change in amplitude is most often attributed to the automatic gain controller or bad contact between the sensor and the skin. Even though the raw signal amplitude does contain useful information regarding the current state of the cardiovascular system, the amplitude of the signals obtained via commercial devices should not be used as a feature [34].

Recently, some effort has been made to normalize the PPG signal across patients, even when measured by different devices [16]. The authors reported successful use of the PPG amplitude as a feature for BP estimation, however, several assumptions were made and the limitations were highlighted.

Accordingly, we have decided to omit any amplitude-related features from our work, since we are dealing with two datasets originating from different devices. The first was recorded using different hospital devices and the second was recorded using a wristband, however the aim is to have a general system applicable to any PPG, independent of the recording device.

3.2.2 Feature extraction

PPG features commonly describe the morphology of an individual cycle in the time domain, using durations of certain characteristic shapes within the PPG waveform, or areas above and under certain parts of the waveform. A large number of time-domain features, which include the mentioned durations and areas of the waveform, were initially computed in accordance with the related work [12, 13, 14, 15]. These features are given in Table 3.3 and shown on an example PPG segment in Figure 3.11.

Feature	**Description**
T_c	Cycle duration
T_s	Time from start of cycle to systolic peak
T_d	Time from systolic peak to end of cycle
T_{nt}	Time from systolic peak to diastolic rise
T_{tn}	Time from diastolic rise to end of cycle
S_1	Area under the curve (AUC) from start of cycle to max upslope point
S_2	AUC from max upslope point to systolic peak
S_3	AUC from systolic peak to diastolic rise
S_4	AUC from diastolic rise to end of cycle
AUC_{sys}	$S_1 + S_2$
AAC_{sys}	Area above the curve (AAC) from start of cycle to systolic peak
AUC_{dia}	$S_3 + S_4$
AAC_{dia}	AAC from systolic peak to end of cycle

Table 3.3: Elaborations of the time-domain features that were used and are shown in Figure 3.11.

In addition to the features describing the morphology of the PPG waveform, the following features describing the state of the cardiovascular system were computed:

1. *AI - Augmentation Index*: a measure of wave reflection on the arteries.

$$AI = \frac{diastolic\ rise\ amplitude}{systolic\ peak\ amplitude} \tag{3.10}$$

2. *LASI - Large Artery Stiffness Index*: an indicator of arterial stiffness, which is denoted as T_{nt} in Table 3.3 and Figure 3.11.

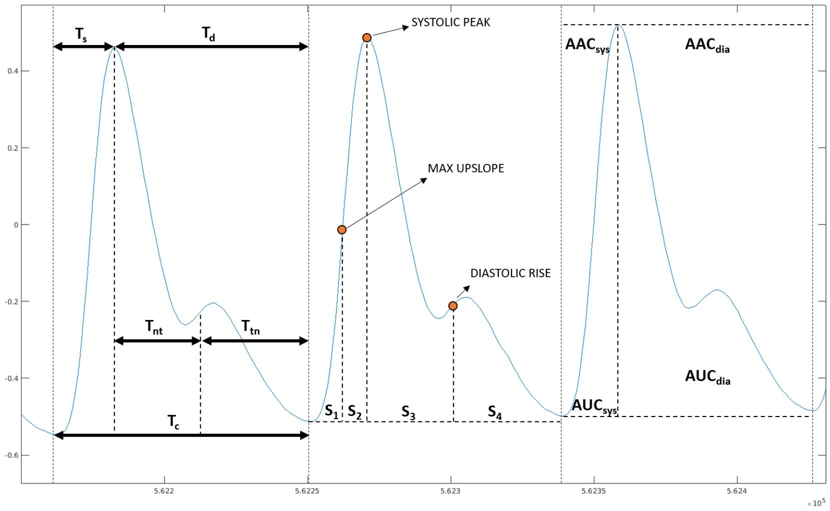


Figure 3.11: Time domain features describing the morphology of the PPG signal on a per-cycle basis. The features are described in more detail in Table 3.3.

The set of features was further expanded with additional features from the complexity-analysis [41] and frequency [16] domains.

1. *Complexity analysis features*: signal complexity and mobility are computed for the 30-second PPG segment containing the current cycle as

$$E_0 = \sqrt{\frac{\sum_{i=1}^{N} PPG(i)^2}{N}}, \tag{3.11}$$

$$E_1 = \sqrt{\frac{\sum_{j=2}^{N-1} d_j^2}{N-1}}, \tag{3.12}$$

$$E_2 = \sqrt{\frac{\sum_{k=3}^{N-2} g_k^2}{N-2}}, \tag{3.13}$$

where PPG is the PPG signal, d is the first order derivative of x and g is the second order derivative of x.

$$Complexity = \sqrt{\frac{E_2^2}{E_1^2} - \frac{E_1^2}{E_0^2}}, \tag{3.14}$$

$$Mobility = \frac{E_1}{E_0} \tag{3.15}$$

2. *Frequency domain features*: amplitudes and phases of the frequency-domain representation of the 30 second PPG segment containing the current cycle, which is given as

$$PPG_{\text{FREQ}} = \text{fft}(PPG_{\text{TIME}}), \tag{3.16}$$

where PPG_{FREQ} is the frequency domain representation of the 30-second PPG segment from the time domain (PPG_{TIME}). The amplitudes are obtained as

$$|PPG_{\text{FREQ}}[k]| = \sqrt{(Re(PPG_{\text{FREQ}}[k]))^2 + (Im(PPG_{\text{FREQ}}[k]))^2}, \tag{3.17}$$

and the phases are computed as

$$\angle PPG_{FREQ}[k] = arctan\left(\frac{Im(PPG_{\text{FREQ}}[k])}{Re(PPG_{\text{FREQ}}[k])}\right), \tag{3.18}$$

where $Re(PPG_{\text{FREQ}})$ is the real and $Im(PPG_{\text{FREQ}})$ the imaginary part of the frequency components of the PPG after FFT.

3.2.3 Feature selection

In total, 46 features were considered. Some of them may be extremely important and might possess vital information about the BP, while others might be insignificant, having little to no information about the BP. The importance of the features was analyzed and features with non-zero importance were selected. This allows us to determine the highest quality features and decreases the computational complexity with removal of irrelevant features.

The RReliefF algorithm [42] was chosen for feature selection. It is a modification of the ReliefF algorithm, suitable for regression problems with continuous target variables.

The original ReliefF algorithm attempts to estimate the importance of each feature according to how well its values distinguish between instances that are near to each other. It does this by finding k nearest neighbours to a given instance from the same class (nearest hits) and also k nearest neighbours from each of the different classes (nearest misses). Subsequently, it updates the quality estimation for a given feature depending on its ability to separate the hits and misses. In the RReliefF modification, which can be used for regression, the class information is replaced by a kind of probability that two instances are different. Details of the algorithm are disccussed by Robnik Šikonja and Kononenko [42].

The advantage of the RReliefF algorithm compared to other feature selection algorithms is two-fold. First, the algorithm can easily be used for a regression problem, as discussed earlier. Second, the result is very intuitive, as it gives the ranking of importance for all the considered features.

We applied the algorithm to a subset of 10% of all data chosen randomly. This was repeated 10 times. All the features with non-zero relevance, as computed by the algorithm, were considered in each iteration, and their importance was saved. Finally, the average importance of each feature was computed and evaluated. The average importance of the features as given by RReliefF algorithm is shown in Figure 3.12.

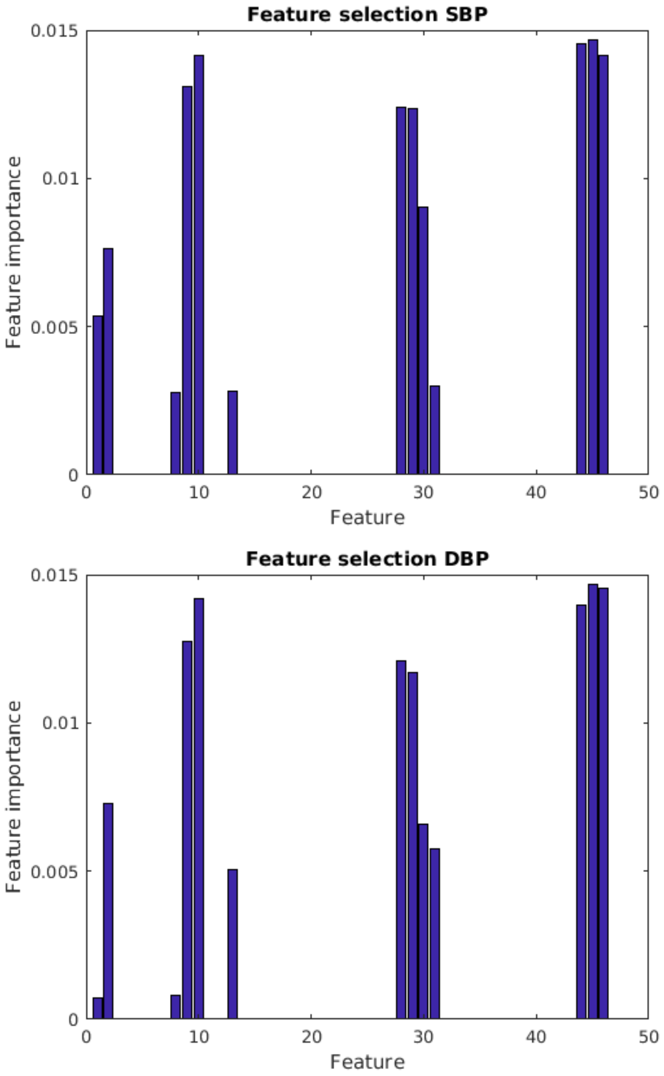


Figure 3.12: The output of RReliefF algorithm, which shows the feature importance for each of the considered features. Both subplots show the same feature space.

We can see that the same features were given non-zero importance for both SBP and DBP. Thus, we have used this subset of features in all subsequent machine learning experiments.

3.3 Machine learning

Since we are trying to estimate BP, which is not a discrete but a continuous target variable, we are dealing with a regression problem. In regression analysis we attempt to determine a functional relationship between the continuous dependent variable (BP) and independent variables (PPG features). Such a relationship may be simple (e.g., linear) or complex (e.g., non-linear) [43].

3.3.1 Classical regression

The regressive nature of the problem was briefly highlighted at the beginning of this section. Regression models involve:

- the unknown parameters of the model, denoted as $\boldsymbol{\alpha}$,
- the independent variables $\mathbf{X}$ and
- the dependent variable $\mathbf{Y}$.

An algorithm then finds a function f that relates the $\mathbf{X}$ and $\mathbf{Y}$ as well as possible, given a certain criterion. It can be formulated as

$$\mathbf{Y} \approx f(\mathbf{X}, \boldsymbol{\alpha}) \tag{3.19}$$

In our domain, the vector $\mathbf{X}$ comprises the computed features, which describe the PPG. The dependent variable $\mathbf{Y}$ is either the SBP or the DBP, and the parameters $\boldsymbol{\alpha}$ are algorithm-specific. The MAE is our chosen criterion, as given in Equation 3.20.

We chose to use the expression "classical regression" exclusively to separate these algorithms from the deep learning neural network regression.

As there is no consensus on a single best algorithm for such a problem, we have evaluated a number of them:

- **Linear regression** – A linear approach for modeling the function f. It attempts to model the given training data points with a linear function, which corresponds to fitting a line to the given set of known data points. The line is fitted using the least squares approach. An overdetermined set of linear equations is solved, by using solution estimates that produce the lowest sum of squared errors (SSE) [43].

- **Regression trees** – Regression trees are a non-linear approach, as they first recursively partition the space of data points into smaller regions, and then fit a simple constant model (e.g., mean, median, etc.) in each of the partitions. The partitioning of the data in terms of minimizing the SSE across all potential splits is computationally infeasible, thus a greedy method is commonly used [44, 45].

- **Model trees** – Similar to regression trees, with the notable difference being that the fitted model in each of the partitions is a function (e.g., a linear function) and not a constant value.

- **Ensembles of trees** – Trees can be merged into ensembles of trees, which often show superior performance compared to individual trees. The ensemble is based on either bootstrap aggregation (bagging) or boosting. Bagging fits the trees many times to different bootstrap-sampled data, thus creating slightly different trees each time. The bagged prediction is the average prediction from these trees [45]. Boosting, on the other hand, creates a default tree on all the data and then creates additional trees focusing on the data that the default tree predicted with the greatest error. The final prediction is given as a weighted average of the individual tree predictions [45].

- **Random Forest** – Similiar to the ensembles, with the notable difference that the algorithm creates trees using a different set of features for splitting each time. This creates much more varied trees which are not very correlated. The final prediction is again given as the average of the individual tree predictions [45].

There are other well established regression methods, such as Support Vector Machines (SVM). Due to a relatively large dataset and limitations in time and computational power, only a subset of possible regression algorithms was evaluated.

3.3.2 Symbolic regression and genetic programming

Symbolic regression [20] is a type of regression analysis that attempts to create mathematical expressions that best fit a certain dataset. It is closely related to genetic algorithms [46, 47], as it uses the same concepts of evolving a population using cross-over and mutation.

The idea is to first create random mathematical expressions from a pre-determined set of building blocks, which include elementary mathematical operators and numerical operands called terminals. These terminals are represented by random constants or the pre-computed PPG features. The initial random expressions represent the initial population, and are explicit equations for predicting SBP and DBP. In the following iterations, the expressions or individuals of the population are evolved using typical genetic programming operators, namely selection, crossover and mutation. This improves the equations to more accurately predict the BP. The selection of best individuals is based on the fitness function, which is chosen to be MAE.

G. Sannino et al. [48] claim to be the first to adopt such an approach for BP estimation. Their proposed approach has been modified and expanded, as described in the following sections.

3.3.2.1 Individuals and population

The purpose of an individual is to be a bridge between the real-life problem context and the problem-solving evolutionary space [47]. Thus, an individual was defined as a mathematical expression or function, created from a pre-determined set of building blocks, which represents an explicit formula for computing the SBP or DBP. Each individual can be represented as a tree structure, consisting of varying number and types of building blocks, as shown in Figure 3.13. The tree nodes correspond to the operators, the sub-trees to the operands, and the tree leaves to the terminals. The depth of the tree is limited in advance, as we do not wish to consider expressions beyond a certain complexity.

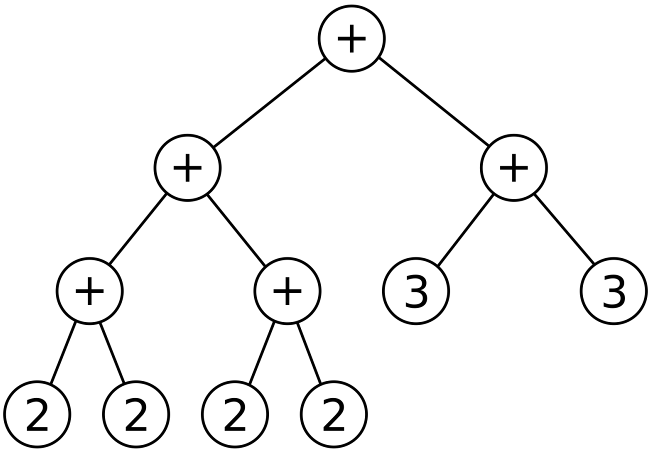


Figure 3.13: The expression $((2 + 2) + (2 + 2)) + (3 + 3)$ shown as an expression tree.

Each feature described in Section 3.2.2, alongside a set of elementary mathematical functions and random constants, comprised the set of the building blocks for the expression trees. The full set is given in in Table 3.4.

Some operators take a single input (e.g., sine), while others take two inputs (e.g., addition). Thus, the individuals are not strictly binary expression trees, since each node can have either one or two children, depending on the arity of the operator that is represented by the node.

The purpose of the population is to represent different possible solutions. While the individuals are static and do not change, the population does, as it is a unit of evolution [47]. A relatively large fixed number of randomly created individuals comprises the initial population, which is then evolved through generations, by using biology-inspired genetic operators, which attempt to improve each successive generation according to a chosen fitness function.

Description	Type	Symbol	Arity (Nr. of operands)
Constant	Operand	C	0
Feature	Operand	x_i	0
Addition	Operator	$+$	2
Subtraction	Operator	$-$	2
Multiplication	Operator	$*$	2
Division	Operator	$/$	2
Sine	Operator	sin	1
Cosine	Operator	cos	1
Natural logarithm	Operator	ln	1
Natural exponentiation	Operator	e^{x_i}	1
Square root	Operator	$\sqrt{x_i}$	1
Absolute value	Operator	$\lvert x_i \rvert$	1

Table 3.4: Potential building blocks of each individual (mathematical expression) in the population.

3.3.2.2 Fitness function

The role of the fitness function is to guide the evolution, by representing the requirements that must be adapted to. It assigns a quality measure to individuals and thus defines what improvement is in the context of evolution [47].

In our problem domain, we would like each expression with PPG input variables to return an accurate prediction about the SBP and DBP at those PPG values. It thus makes sense to use the MAE as the fitness function, since it describes the quality of the solution well. It tells us how well the predictions match the observed values and is defined as

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - \lambda(x_i)|}{N} = \frac{\sum_{i=1}^{n} |e_i|}{N}, \tag{3.20}$$

where y_i is the observed value of BP and $\lambda(x_i)$ is the model-predicted value for the instance x_i [30]. Their difference is the error, denoted as e_i and N is the total number of predictions.

A feasible secondary fitness function could be the complexity or size of the expression tree.

3.3.2.3 Selection mechanism and genetic operators

Genetic variation is the basis of evolution. The role of genetic operators is to create new individuals from those chosen by the selection mechanism [47]. The mentioned mechanism and operators are chosen as follows:

- **Selection** – It governs the reproduction process, as it chooses the specific individuals that will undergo said reproduction. It should favor high-quality individuals in order to create potentially superior offspring. Low-quality individuals must not be discarded, but rather given a low selection chance. This is achieved by using the *tournament* selection mechanism, which first chooses a fixed number of individuals at random (with replacement) and then selects the best among them, with respect to the fitness function, for reproduction. The expression trees corresponding to equations with low MAE are expected to be chosen more often.

- **Crossover** – It merges the information from two parent individuals into two offspring. A subtree is randomly selected in each parent tree and then these subtrees are swapped. The depth of the child trees must not exceed the pre-determined maximum depth. If a child tree is too deep, it is discarded and one of the parents is chosen and copied into the new population.

- **Mutation** – It encourages genetic diversity and is always a stochastic process. A random inner node is chosen in a child tree and its operator is changed to another random operator with the same arity. Constants in the leaves of the trees may also be mutated.

- **Copy** – A small percentage of parent individuals are simply copied into the next generation. The *elitism* concept was followed, which means that the best individuals in a population are always copied without mutation. This ensures that the new population will be at least as good as the previous, but improvements are expected.

These operators are applied until a new generation with the same number of individuals is created. Since elitism was implemented, the convergence of candidate solutions towards a good solution is expected [48].

An initial generation of 100 solutions (equations) is first created randomly. Solutions are evolved until the best individual meets a required quality threshold or until 1000 subsequent generations show no notable improvement regarding the fitness function (MAE decrease greater than 1 mmHg). The size of each solution is limited to 100 building blocks. When the process stops, the best individuals created during the evolution are considered.

3.3.3 Deep learning

In recent years, deep learning has arisen as a cutting-edge approach to machine learning. There are numerous fields at which deep learning shines, including image recognition, music recognition, disease recognition, games, etc. In 2017, Google's AlphaGo [49] defeated the world champion in the hugely complex game of Go in a best-of-three series, which marked the most recent breakthrough in AI surpassing humans in activities that were previously considered a human domain. The approach is not new, as it is based on artificial neural networks (ANNs), which have been around for decades.

ANNs are inspired by biological neural networks and comprise of interconnected assembly of simple elements called nodes or neurons. An early form is the multilayer perceptron (MLP), which commonly has a small number of neurons and all the connections are pointing from the input towards the output, meaning the perceptron is always feed-forward. The neuron is a processing element containing an activation function. This activation function takes n input connections. Each input connection has a corresponding training weight w_i. The neuron's activation function returns an output connection, computed with the activation function. Commonly used activation functions are the sigmoid, tangens hyperbolicus and rectifier (ReLU) [21].

The training is commonly done using the gradient descent optimization algorithm and backpropagation. These are used to adjust the training weights W based on the error calculated at the output. This error is then distributed back through the network layers and the weights are adjusted so that the error decreases in the subsequent iterations.

Stacking a number of neurons into connected layers gives rise to ANNs. If the number of layers between the input and the output layer is large enough (≥ 2), the network is considered deep, as shown in Figure 3.14.

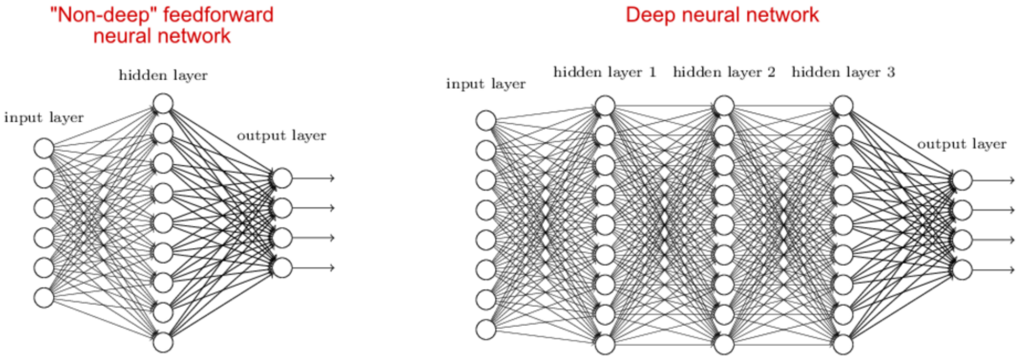


Figure 3.14: A schematic example of "non-deep" and deep ANNs [50].

With the large increase in size of training datasets and increase in computational power through parallelization, deep ANNs have reached their potential. Since an extremely large number of weights must be updated during training, this was computationally infeasible in the past, but became quite possible with the appearance of powerful graphics processing units (GPUs), which allow huge parallelization. Nowadays, modern deep neural networks can have up to 1000 layers and millions to billions of neurons, making them capable of modelling extremely complex non-linear relationships [21].

A vital aspect that differentiates deep ANNs from traditional machine learning methods is the fact that ANNs are capable of deriving features on their own, from raw input data. In the past, creating features was typically in the domain of the researcher, who then fed these features into the learning algorithm. Deep ANNs allow researchers to simply feed them large amounts of labeled input data, and then the ANNs can derive the features and do the learning on their own. Despite this, quality hand-crafted features can improve the performance of the model. ANNs can also take a combination of raw signal and hand-crafted features, making them even more powerful.

Another advantage is the fact that deep ANNs are capable of modelling extremely complex non-linear relationships, thus having more expressive power compared to some traditional methods.

We have considered using both cleaned raw PPG signal and hand-crafted features individually, as well as a combination of both. The requirement for the raw input data is that each instance has the same length. This is easy to ensure when using features, however it is more difficult when using raw PPG cycles. We have resolved this by centering each PPG cycle in a 2-second window, which is long enough to always contain a single beat, as seen in Figure 3.15.

Despite the mentioned advantages, deep ANNs cannot be very effectively used out of the box. Besides the high computational complexity, a major challenge lies in determining optimal parameters for a network to be able to learn effectively. The topology of the neural network is the first unknown, as the network can be either deep or shallow, wide or narrow. The common agreement is that deeper is better. However, this is not always true and should not be taken as a fact. Additionally, the computational complexity increases with greater depth. The next unknown is the type of activation function to use, and the type of optimizer algorithm, which will ensure the convergence to local minima. The types of layers is the next parameter of interest, as there are different types of layers (fully connected, convolutional, recurrent, Long-Short-Term-Memory, etc.), suitable for different problems. Finally, a large number of other parameters must be chosen, such as the

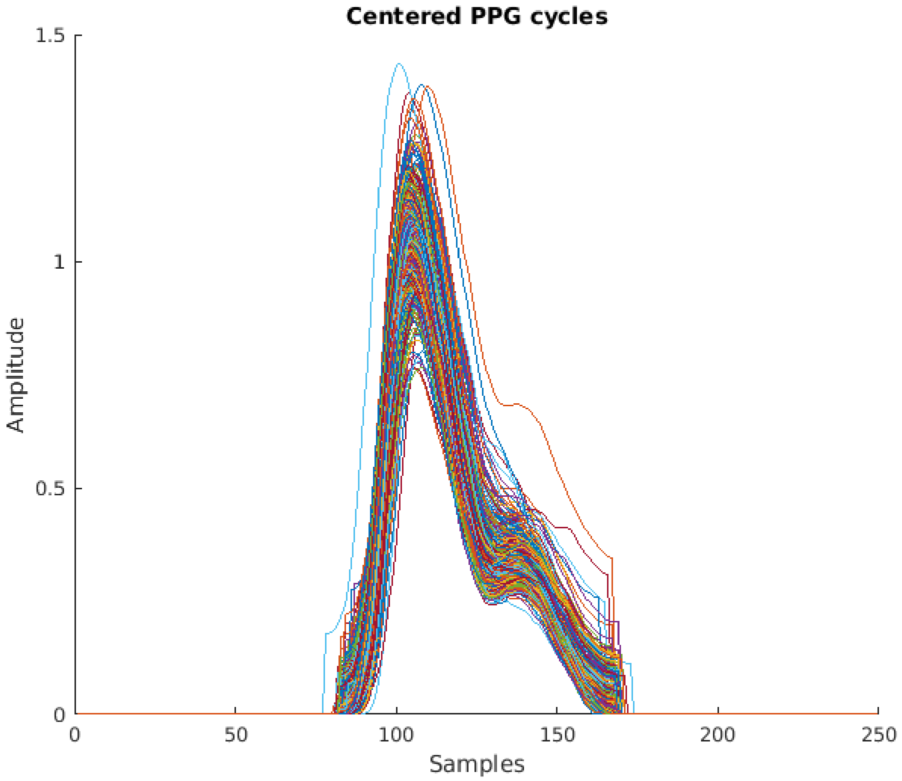


Figure 3.15: 500 PPG cycles, where each is centered in a 2-second window.

learning rate, potential dropout and regularization parameters, which attempt to ensure that the network will not overfit to the training data [21].

As mentioned, the input layer to our network received either the computed features, the centered high-quality raw PPG cycles or a combination of both. The output layer of the network always consisted of two neurons, one for SBP and the other for DBP. The chosen metric to judge the performance of the model was MAE. Some additional parameter choices regarding the topology, learning rate and types of layers were explored, and are detailed the following chapter.

Chapter 4

Materials

Two distinct datasets were used in our work. The clinical dataset was larger and was collected from the treated patients in a hospital environment, while the everyday-life dataset was smaller and was collected at the Jožef Stefan Institute (JSI), during everyday activities of the employees.

4.1 Clinical dataset description

The MIMIC database is available at `https://physionet.org/cgi-bin/atm/ATM` and is widely used in experiments and competitions dealing with bio-medical signals [22]. The original version of the database contains recordings of 72 patients, who were treated in a hospital. During their treatment, a variety of vital functions were monitored, and a number of signals were recorded simultaneously for substantial periods of time, typically for several hours. These include, among others, PPG and BP, making the dataset useful for our experiments.

The data was collected using hospital equipment, which means that the ABP was measured invasively and the PPG was most likely recorded using the fingertip PPG sensor commonly found in hospital settings. ground-truth BP is thus measured using the most precise possible measurement.

The range of BP values in the raw data is extreme, ranging from 0 mmHg to over 1000 mmHg. Such extreme values are not relevant for our work and were thus removed during the signal preprocessing procedure described in Chapter 3. After the preprocessing, the distribution of BP values is shown Figure 4.1.

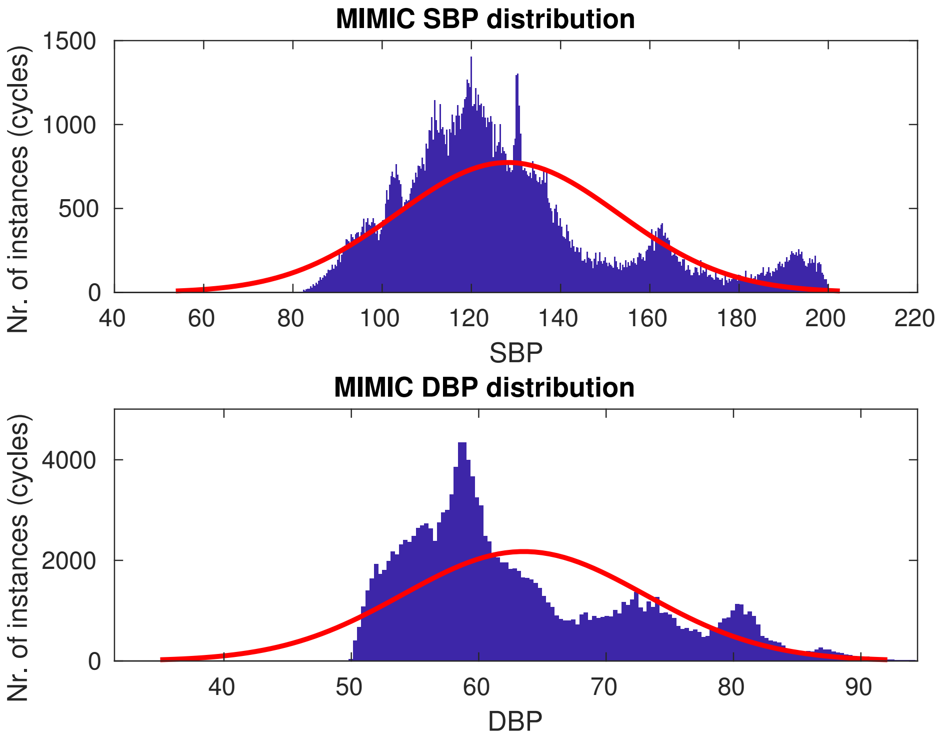


Figure 4.1: Distributions of SBP and DBP instances for the clinical dataset. Red lines show the normal distribution.

Some descriptive statistics about the clinical dataset are given in Table 4.1.

The details of the collection procedure are unknown, however it is most likely that the clinical measurement protocol was followed, since the data was obtained from hospitalized patients [2]. All the patients are anonymous and no detailed information about them is known (e.g., their age, sex, etc.).

The recording for each patient is typically continuous and lasts for a

	SBP [mmHg]	**DBP [mmHg]**
Maximum	199	93
Minimum	78	49
Mean	128	63
Standard deviation	24	9

Table 4.1: Descriptive statistics about the clinical dataset.

few hours. There is no general set of signals which was measured for each patient. This is evident by the fact that different patients have different sets of signals available in the database (e.g. some patients do not have the PPG signal recorded, but might have ECG and ABP).

All the patients in the MIMIC database with both PPG and ABP signal were initially considered. The data was first fed into the pre-processing module. Some of the patients were discarded due to major anomalies in either PPG or ABP signal. In case less than a minimum threshold of 10 minutes of quality data remained for a given patient after the SQI cleaning, that patient was discarded and that data was not used in the experiments.

After the preprocessing was complete, 41 patients had enough high-quality data remaining. For those patients who had more than 1 hour of high-quality signals, data was subsampled by uniformly taking 20 3-minute segments. Additionally, the SBP and DPB value within each 3-minute segment was set to be the mean of the cycle based BP values for all the cycles within this segment. This was done with the purpose of simulating an everyday life setting, in which BP is not known on a per-cycle basis, and does also not change abruptly.

The final post-preprocessing clinical dataset totalled at around 160 000 instances (roughly around 30 hours of signals), where one instance corresponds to one PPG cycle. For these, features were calculated and selected as described in Chapter 3. Additionally, for the purpose of deep learning, raw cycles were saved to be used with a convolutional neural network.

4.2 Everyday-life dataset description

The second dataset was collected at JSI using the Empatica E4 wristband for the PPG and a digital cuff-based Omron BP monitoring device for the ground-truth BP, as is common in such experimental settings in related work. The collection procedure was conducted in accordance with the standardized clinical protocol [2]. In an ideal situation, the ground-truth BP should be measured as ABP within an artery, but due to the invasive nature of such ABP measurement, this is not feasible in an everyday-life situation, so the digital cuff-based monitor was used as a good replacement. An upper-arm cuff-based monitor was chosen over a wrist-based one, as the latter is less accurate and extremely sensitive to body position.

The BP measurements were done periodically during the subjects' daily routine. The subjects were encouraged to measure their BP at least once every 30 minutes or more often, however, no restrictions were forced upon them, allowing them to make measurements more or less often, depending on their daily schedule.

In the first completed phase of the data collection, 10 healthy subjects were considered, 7 male and 3 female. Only parts of the PPG signal 3 minutes before and after each BP measurement point were taken into consideration, as the measured BP value is only relevant for a short time. Ideally, the BP would be measured more often, however, this would place further stress on the subjects and was not possible during their everyday routine. Also, taking less than 3 minutes of signal before and after the measurement would be preferred, however, due to limited amount of data (5-10 measurements per day) that would lead to insufficient amount of data for the experiments.

The same procedure as was used with the clinical dataset was once again followed, starting with the signal preprocessing. The effects of the preprocessing on the everyday-life dataset were more substantial compared to the clinical dataset, as the PPG signal collected with a wristband during daily routine contains notably more artefacts and erratic amplitude variations compared to the hospital-collected PPG. Subsequently, two subjects were discarded during the preprocessing due to an extremely small amount of data remaining after the preprocessing (originally very short recordings with only a couple of BP measurements). The distribution of BP values is shown in Figure 4.2.

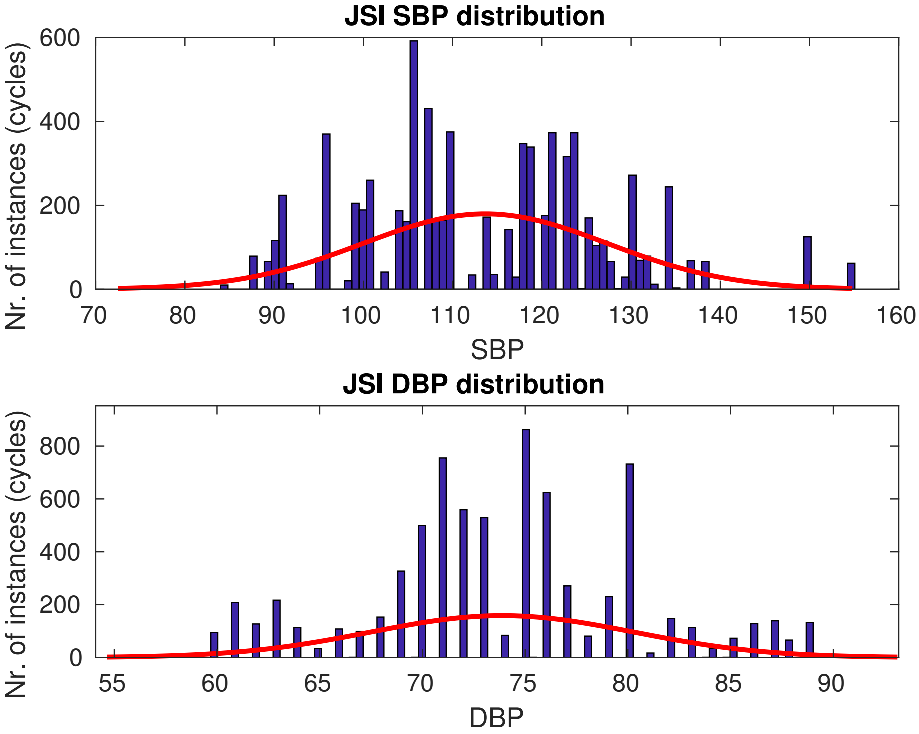


Figure 4.2: Distributions of SBP and DBP instances for the everyday-life dataset. Red lines show the normal distribution.

Some descriptive statistics about the everyday-life dataset are given in Table 4.2.

	SBP [mmHg]	**DBP [mmHg]**
Maximum	155	89
Minimum	84	60
Mean	114	74
Standard deviation	14	6

Table 4.2: Descriptive statistics about the everyday-life dataset.

No subsampling was done on the everyday-life dataset, as the amount of data is much lower compared to the clinical dataset. Finally, around 10000 instances remained, corresponding to roughly 3 hours of signal.

Chapter 5

Experiments and evaluation

In order to evaluate the performance of the proposed system, several experiments were designed, using both datasets described in Chapter 4.

5.1 Experimental setup

Three generic experiments were designed and a special experiment was created for the symbolic regression. Only a suitable subset of all experiments were conducted for each dataset. The whole set of experiments was defined as follows:

1. **Experiment 1 (Symbolic regression)** – Data instances of all subjects were first merged. They were then split into training (66%) and testing (34%) set. Instances were not shuffled. Evaluation of each generation in respect to the fitness function (MAE) was continuously conducted on the testing set.

2. **Experiment 2** – Data instances from each individual subject were again split into training (66%) and testing (34%) set. Instances were not shuffled. Evaluation was conducted per-subject, and the average of the MAEs across all subjects was considered as the final metric.

3. **Experiment 3** – Data instances of all subjects were first merged and

shuffled. Subsequently, k-fold cross-validation was conducted, where the instances were first split into k folds of nearly equal size. Then $k - 1$ parts of the data were used for training while 1 fold was used for testing. This was repeated k times, with a different fold used for testing each time. The average MAE across all folds was used as the final metric.

4. **Experiment 4** – Leave one subject out (LOSO) evaluation, where all but one subject were used for training and the remaining subject was used for testing. This was repeated until each subject was used for testing once. The overall average MAE across all iterations was again computed as the final metric.

In all experiments we ensured that there is no overlap between the training and testing set, meaning no same instance ever appeared in both sets.

Experiment 1 was designed as a simplified version of the LOSO experiment, due to high computational power requirements and time consumption of the LOSO experiment in combination with the symbolic regression algorithm. Experiments 2-4 are generic and were used for all the other algorithms.

MAE [30] was used as the evaluation metric in all of the experiments, as it is suitable to describe the performance of the model and also widely used in related work. ABP was taken as the ground-truth for clinical dataset while the Omron digital monitor measurements were used as ground-truth for the everyday-life dataset. The predictive performance was always compared to a dummy regressor, which predicted the mean SBP or DBP value.

As a large number of different algorithms was evaluated, not all the experiments were done for each algorithm. Some algorithms are more suitable for certain types of experiments (e.g., deep learning should not be used on small amounts of data, which is the case in experiment 2).

5.1.1 Personalization of the models

The LOSO experiment best validates the robustness and generalization performance of a model. As it initially yielded poor results for all the algorithms, while k-fold cross-validation (experiment 3) and per-patient validation (experiment 2) performed well, we considered personalization of the general model for each patient in the LOSO experiment (experiment 4).

We have considered using a small amount of each patient's data for supervised training, as experiments 2 and 3 have shown that personalized models achieve low average MAE. This most likely happens due to each patient having a subtly unique cardiovascular dynamic and relation between PPG and BP. This assumption was additionally confirmed by doing cycle morphology analysis, during which it was established that similar cycle shapes do not necessarily signify similar BP values. Due to the mentioned factors, personalization of the trained models was considered in an attempt to improve the predictive performance of the general model.

The regression models in the LOSO experiment were again trained using all the subjects except the left out. This time, however, the models were personalized using some instances from the left out subject. The instances of the left out subject were grouped by their BP values. These groups were then sorted from lowest to highest BP. Afterwards, every n-th group ($n = 2, 3, 4, 5, 6$) of the instances was taken from the testing data and used in training in order to personalize the model for the current patient. This ensured personalization with different BP values, as taking just a single group of instances, or several groups in sequence, gives little information, since the BP is the same or similar for all of them.

Minimum personalization was 0, while maximum personalization was roughly 50%, corresponding to 30 minutes of signal or less. In practice, this would require the user to make 5-10 BP measurements, depending on the amount of PPG signal around the measurement, that we consider having this same BP value. A small number of BP measurements alongside the corresponding PPG signal allows for the maximum personalization of the predictive model. This means that a user obtains a higher-quality personalized model, compared to the initial general model, within a day, assuming he inputs a few ground-truth BP measurements manually.

5.1.2 Algorithm details

Each of the considered algorithms comes with its own set of parameters or hyper-parameters, which can potentially be fine-tuned for optimal performance. Due to the large amount of data instances and extensive experimental setup, we used the default set of hyper-parameters for each of the classical regression algorithms. These algorithms were mostly used as given by the out of the box implementations in MATLAB [51]. This is a potential point for improvement, which will be discussed in the conclusion.

For deep learning, we have used the Keras Python library [52], which runs on top of the TensorFlow framework [53].

Hyper-parameter optimization is an extremely important open problem in deep learning, which does not have an elegant solution. Each of the hyper-parameters for deep learning can profoundly influence the network and its performance. These are often set experimentally, or in best case using a random search or grid search hyper-parameter optimization approach, which can require extreme amounts of time and computational power [54].

We have conducted a very limited grid search type of hyper-parameter optimization, by exploring some high-level options (e.g., shallow vs. deep, wide vs. narrow, etc.), as is common practice when using deep learning on a new problem. The hyper-parameters and their explored values are given in Table 5.1.

After doing experiments with the given values, we have come to a best-performing neural network with the hyper-parameters marked in bold in Table 5.1. A deeper network (4 hidden layers with 512–256–128–64 neurons) proved best, while the learning rate was set to the rather low value of 0.0001, as higher learning rates caused erratic movement of the loss function. ReLU activation function and Adam optimizer have shown the best performance.

Hyper-parameter	**Explored values**
Topology of the network	{shallow, **deep**, wide, narrow}
Activation function	{**ReLU**, sigmoid, tangens hyperbolicus}
Learning rate	{0.1, 0.01, 0.001, **0.0001**, 0.00001}
Types of layers	{**fully-connected**, convolutional}
Dropout rate	{0, **0.25**, 0.5}
Number of training epochs	{10, 25, **50**, 100}

Table 5.1: Explored hyper-parameter values for the deep learning regression. Best-performing hyper-parameters are marked in bold.

Adam was used throughout the experiments as its performance is typically very good on a wide range of problem domains [55]. A 0.25 dropout rate was used to prevent overfitting of the network. Dropout forces the network to drop random connections between neurons in order to prevent fast overfitting. The training was done for 50 epochs. It was empirically determined, that the lowest MAE loss is always achieved between 30th and 50th epoch, while additional epochs only require additional time, while offering no improvements of the model.

When using convolutional layers, the input data was set to be the features alongside the raw PPG cycles, centered in a 2-second window. The purpose of the convolutional layer is to derive some features from the signal on its own, while also being capable of keeping relevant hand-crafted features from the start. The predictive performance of the convolutional neural network with both the raw signal and the features as input was nearly identical to that of the fully-connected network with only the features as input. Thus, fully-connected network was chosen, as its training is much faster compared to the convolutional network. We have also attempted to use convolutional network with only cleaned raw signal as input, however the performance was slightly worse compared to using both the cleaned raw signal and the hand-crafted features.

5.2 Clinical dataset (MIMIC database)

5.2.1 Experiments and results

5.2.2 Results

The results are presented in accordance with the descriptions of the experiments in Section 5.1. A table showing the average MAE ($\mathrm{MAE}_{\mathrm{SBP}}$ and $\mathrm{MAE}_{\mathrm{DBP}}$) and the corresponding standard deviations ($\mathrm{STD}_{\mathrm{SBP}}$ and $\mathrm{STD}_{\mathrm{DBP}}$) for both SBP and DBP is given first. The errors of the best performing algorithm are then plotted.

5.2.2.1 Experiment 1 (Symbolic regression)

The symbolic regression achieved the lowest $\mathrm{MAE}_{\mathrm{SBP}}$ of 17.24 and $\mathrm{MAE}_{\mathrm{DBP}}$ of 14.94, as seen in Figure 5.1, which shows the MAE decreasing during the evolution. The complexity of the solutions was rather high (50 – 70 building blocks) and the solutions did not use many different features. This algorithm has proven to be the worst, not even surpassing the dummy performance, which achieved $\mathrm{MAE}_{\mathrm{SBP}}$ of 18.44 and $\mathrm{MAE}_{\mathrm{DBP}}$ of 10.77 on the same 66%-34% training-testing data split.

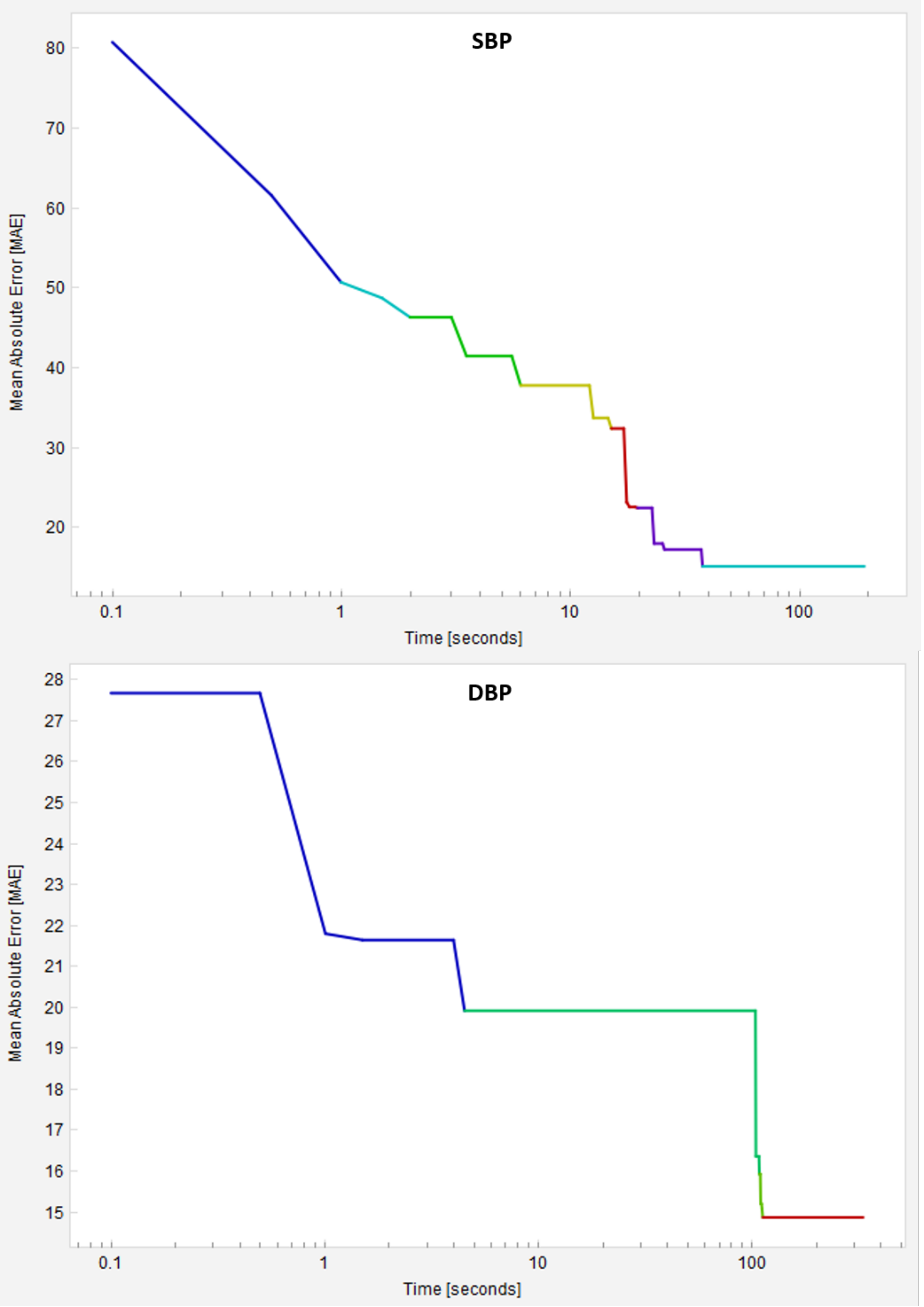


Figure 5.1: Decreasing MAE during the evolution of the symbolic-regression models in experiment 1. Each color corresponds to the complexity of the proposed solution (the number of building blocks of the equation), with green being the simplest, blue being the medium, and red being the most complex among the solutions encountered so far.

5.2.2.2 Experiment 2

When evaluation is done on a per-subject basis, meaning the instances of a subject are split into training and testing set, and the model is built for this subject specifically, the models are completely personalized and thus show optimistic predictive performance compared to a completely general model, as shown in Table 5.2.

	Errors and standard deviations in mmHg			
Algorithm	**$\mathrm{MAE}_{\mathrm{SBP}}$**	**$\mathrm{STD}_{\mathrm{SBP}}$**	**$\mathrm{MAE}_{\mathrm{DBP}}$**	**$\mathrm{STD}_{\mathrm{DBP}}$**
Dummy	10.23	10.72	8.12	7.92
Linear reg.	10.73	10.12	8.49	8.22
Regression tree	8.44	9.42	7.07	7.66
M5 Model tree	10.89	12.34	9.77	8.48
Random Forest	**6.23**	**6.92**	**4.53**	**3.62**

Table 5.2: Average MAE and corresponding STD across all patients in the per-subject evaluation using different regression algorithms. The best result is marked with bold.

The lowest avg. $\mathrm{MAE}_{\mathrm{SBP}}$ is 6.23 mmHg and avg. $\mathrm{MAE}_{\mathrm{DBP}}$ is 4.53 mmHg. Both the $\mathrm{MAE}_{\mathrm{SBP}}$ and $\mathrm{MAE}_{\mathrm{DBP}}$ vary notably between patients, as shown in Figure 5.2. Both lowest avg. errors are achieved using the Random Forest algorithm.

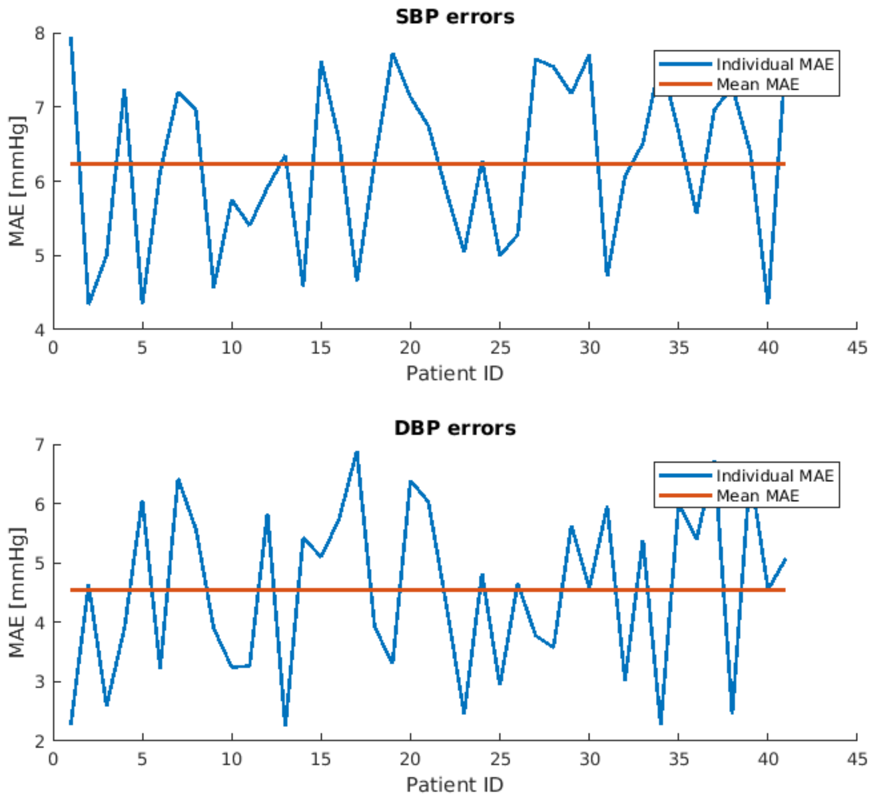


Figure 5.2: Individual MAE per subject and the average MAE across all the subjects using the Random Forest algorithm in the per-subject experiment 2.

5.2.2.3 Experiment 3

In 5-fold cross validation with shuffled instances, partial personalization is implicit, as instances of the same subject are present in the training data due to shuffling. The predictive performance is again optimistic, as shown by the average MAE across all folds for each algorithm in Table 5.3.

The dummy in this experiment performed notably worse compared to experiment 2, as its prediction originated from a different training set. In experiment 2, the dummy always predicted the mean value of the SBP and

	Errors and standard deviations in mmHg			
Algorithm	**MAE$_{\mathbf{SBP}}$**	**STD$_{\mathbf{SBP}}$**	**MAE$_{\mathbf{DBP}}$**	**STD$_{\mathbf{DBP}}$**
Dummy	19.44	16.02	8.53	6.87
Linear reg.	18.47	15.91	8.14	7.98
Regression tree	9.63	9.11	8.47	6.22
Model tree	11.55	11.74	9.98	7.25
Random forest	**7.83**	**7.47**	**3.84**	**3.63**

Table 5.3: Average MAE and corresponding STD across all folds in 5-fold cross-validation using different regression algorithms. The best result is marked with bold. These errors were achieved using the clinical dataset.

DBP within a single patient, while in experiment 3, the dummy always predicted the mean of the training set, which contained instances of all the patients. The BP variations among many patients are much larger than within a single patient, thus making the dummy MAE much larger in experiment 3 compared to experiment 2.

5.2.2.4 Experiment 4

Finally the LOSO experiment was conducted. Each model created in this experiment was first strictly without any personalization. Due to poor performance of such general models, increasing amount of data from the left-out patient was added to the training set in order to personalize the model, as described earlier in Section 5.1. The errors of the models are shown in Figure 5.3.

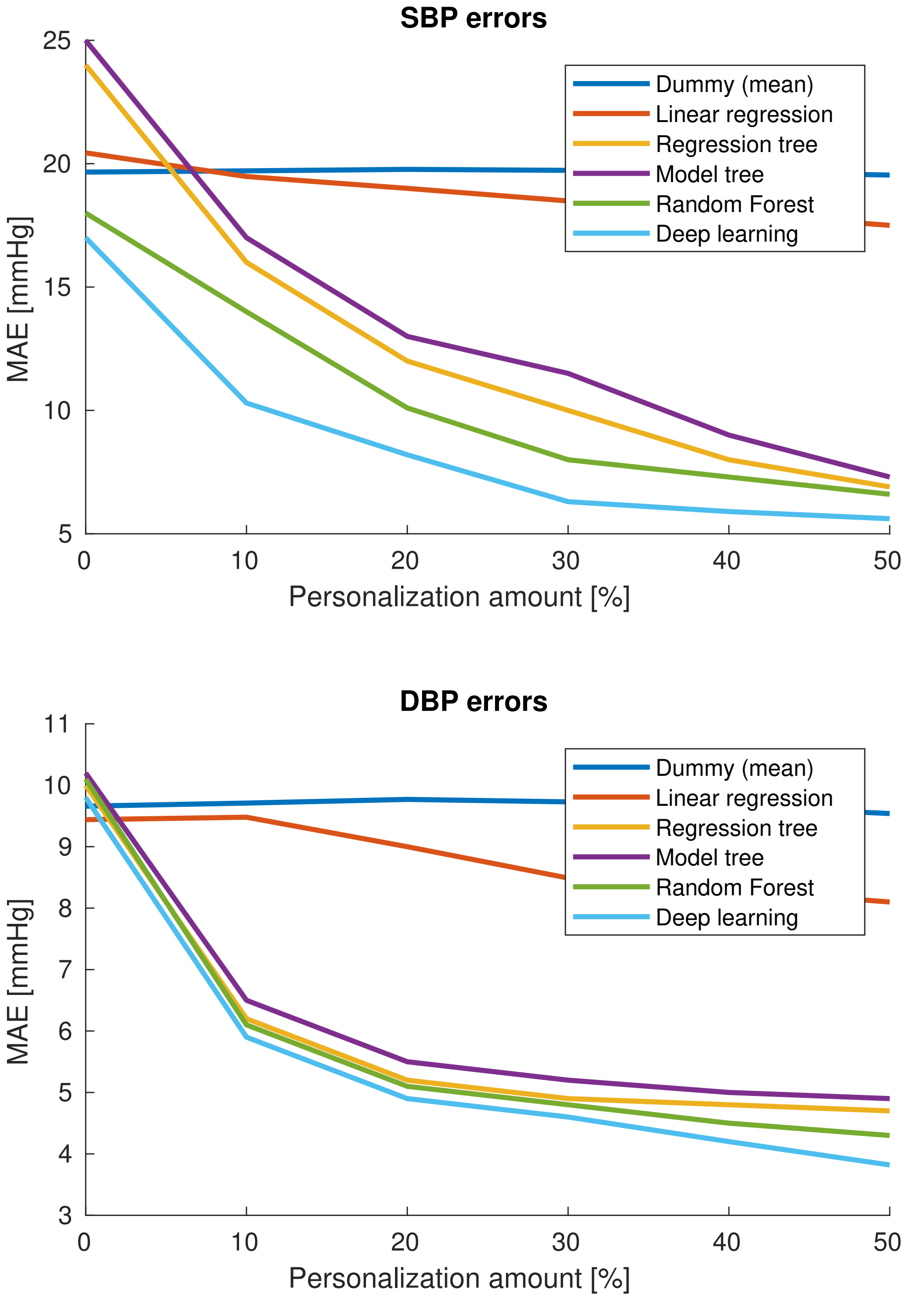


Figure 5.3: Average $\mathrm{MAE}_{\mathrm{SBP}}$ and $\mathrm{MAE}_{\mathrm{DBP}}$ at different amounts of personalization for the clinical dataset in the LOSO evaluation experiment 4.

As expected, both the MAE$_{\text{SBP}}$ and MAE$_{\text{DBP}}$ decrease at increasing amounts of personalization. At 50% personalization, the errors typically approach or surpass those of 5-fold cross-validation. Deep learning achieves the lowest overall avg. MAE with the clinical dataset across all algorithms, as seen in Figure 5.3. MAE$_{\text{SBP}}$ of 5.61 mmHg and MAE$_{\text{DBP}}$ of 3.82 mmHg are achieved, both at maximum personalization.

The comparison of the results of the best performing algorithms between experiments 2, 3 and 4 is given in Table 5.4.

Best in experiment 2				
Algorithm	**MAE$_{\text{SBP}}$**	**STD$_{\text{SBP}}$**	**MAE$_{\text{DBP}}$**	**STD$_{\text{DBP}}$**
Random Forest	6.23	6.92	4.53	3.62
Best in experiment 3				
Algorithm	**MAE$_{\text{SBP}}$**	**STD$_{\text{SBP}}$**	**MAE$_{\text{DBP}}$**	**STD$_{\text{DBP}}$**
Random Forest	7.83	7.47	3.84	3.63
Best in experiment 4				
	MAE$_{\text{SBP}}$	**STD$_{\text{SBP}}$**	**MAE$_{\text{DBP}}$**	**STD$_{\text{DBP}}$**
Algorithm	0% / 50%	0% / 50%	0% / 50%	0% / 50%
Random Forest	18.66 / 6.32	14.16 / 6.02	10.07 / 4.32	6.31 / 3.37
Deep learning	17.12 / 5.61	15.22 / 6.12	9.82 / 3.82	5.42 / 2.91

Table 5.4: Comparison of best performing algorithms and their avg. MAE and STD across experiments 2, 3 and 4 for the clinical dataset. All the reported results are in mmHg and the percentages signify the amount of personalization data that was used.

5.3 Everyday-life dataset

5.3.1 Experiments and results

A subset of experiments described in Section 5.1 was used for the evaluation on the everyday-life dataset. Experiment 1 was discarded, as the perfromance of the symbolic regression was poor. The per-subject validation of experiment 2 was omitted, as there are only a small amount of different BP values within the data of a single subject. Deep learning regression algorithm was also omitted due to the smaller size of the dataset, which is not suitable for deep learning.

The applied restrictions left us with all the classical regression algorithms, which were used in experiments 3 and 4 for the everyday-life dataset. The results are given in the same format as in Section 5.2.2.

5.3.1.1 Experiment 3

As we are dealing with less data and more importantly fewer different BP values in the everyday-life dataset compared to the clinical dataset, the performance of the dummy regressor is notably increased. This is reflected in the errors given in Table 5.5.

The lowest achieved $\mathrm{MAE}_{\mathrm{SBP}}$ was 8.92 mmHg and the lowest $\mathrm{MAE}_{\mathrm{DBP}}$ was 4.27 mmHg. Both were again achieved using the Random Forest algorithm.

5.3.1.2 Experiment 4

Finally, LOSO experiment with personalization was conducted for everyday-life dataset without the deep learning algorithm, due to the aforementioned smaller amount of data. The results are shown in Figure 5.4.

The comparison of algorithms between experiments 2 and 3 for the everyday-life dataset is given in Table 5.6.

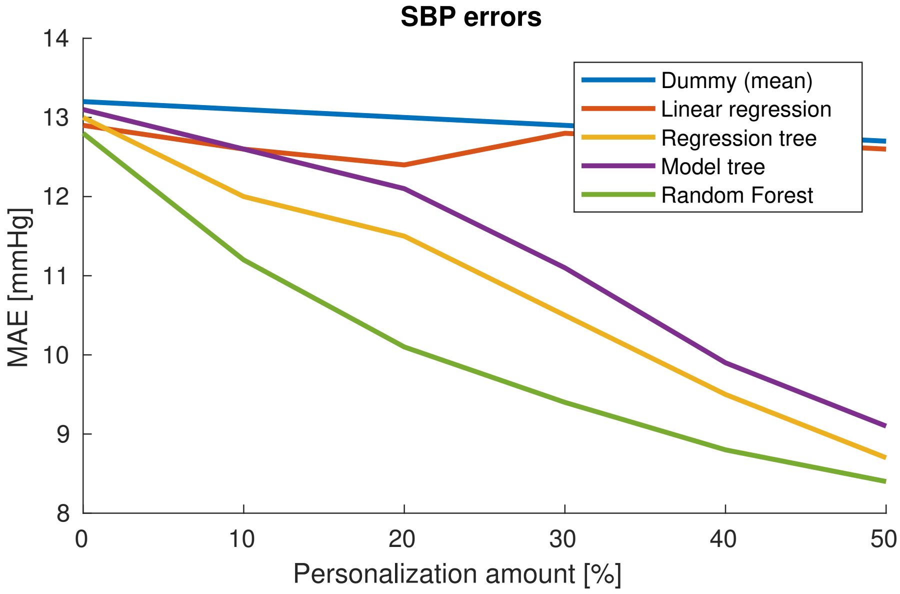
	Errors and standard deviations in mmHg			
Algorithm	**MAE$_{\mathbf{SBP}}$**	**STD$_{\mathbf{SBP}}$**	**MAE$_{\mathbf{DBP}}$**	**STD$_{\mathbf{DBP}}$**
Dummy	11.46	7.51	5.01	3.99
Linear reg.	11.21	8.00	5.01	8.00
Regression tree	9.12	7.90	4.38	3.74
Model tree	10.33	10.01	4.72	3.94
Random Forest	**8.92**	**8.49**	**4.27**	**3.99**

Table 5.5: Average MAE and corresponding STD across all folds in 5-fold cross-validation using different regression algorithms. The best result is marked with bold. These errors were achieved using the everyday-life dataset.

Best in experiment 3				
Algorithm	**MAE$_{\mathbf{SBP}}$**	**STD$_{\mathbf{SBP}}$**	**MAE$_{\mathbf{DBP}}$**	**STD$_{\mathbf{DBP}}$**
Random Forest	8.92	8.49	4.27	3.99
Best in experiment 4				
	MAE$_{\mathbf{SBP}}$	**STD$_{\mathbf{SBP}}$**	**MAE$_{\mathbf{DBP}}$**	**STD$_{\mathbf{DBP}}$**
Algorithm	0% / 50%	0% / 50%	0% / 50%	0% / 50%
Random Forest	12.81 / 8.40	11.03 / 7.53	7.19 / 4.20	5.29 / 3.18

Table 5.6: Comparison of best performing algorithms and their avg. MAE and corresponding STD across experiments 2 and 3 for the everyday-life dataset. All the reported results are in mmHg and the percentages signify the amount of personalization data that was used.

The lowest overall achieved MAE$_{\text{SBP}}$ with the everyday-life dataset was 8.40 mmHg and the lowest MAE$_{\text{DBP}}$ was 4.20 mmHg, both achieved with the Random Forest algorithm at maximum personalization. The comparison between different models in the LOSO experiment is shown in Figure 5.4.



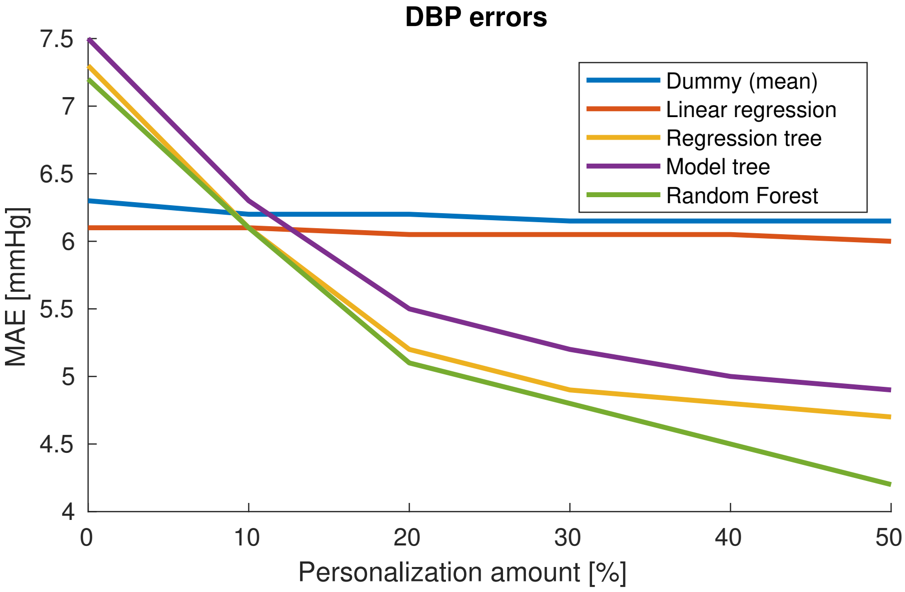


Figure 5.4: Average MAE_{SBP} and MAE_{DBP} at different amounts of personalization for the everyday-life dataset in the LOSO evaluation in experiment 4.

5.4 Discussion of results

It is difficult to compare our results with those of the related work discussed in Chapter 2, since almost every study is dealing with a different dataset and different metrics are used. Most studies used a subset of our clinical dataset, meaning they only used considered a specific parts of a subset of patients from the MIMIC database. Related work also commonly reported ME instead of MAE, which we find inadequate, as it does not fully reflect the performance of the model. Despite this, we believe that our work surpasses the related work, since we achieved comparable errors without some of the dataset limitations commonly imposed in related work, which we highlighted in Chapter 2.

As the different reported metrics and used datasets make the comparison with related work difficult, we will rather focus on requirements imposed by two major standards for BP estimation devices [23].

5.4.1 Comparison with standards

There are two major international standards for BP estimation devices, which are most commonly used to validate BP estimation devices. These two standards propose exact experimental protocol and BP estimation accuracy thresholds that must be met in order for the device to be certified under the given standard [56].

These standards are being followed by the leading digital BP estimation device manufacturers, such as Omron Healthcare. They use these standards and the corresponding validation protocols to validate their devices, which are commonly used in medical institutions around the world [57].

Given the widespread use and medical consensus about the requirements of these standards, we discuss their details and our system in the context of the two standards in the following sections.

5.4.1.1 British Hypertension Society (BHS)

First is the British Hypertension Society (BHS) standard. BHS grades BP measurement devices into three grades, A, B and C, based on their total percentage of absolute errors under three different thresholds, i.e. 5, 10 and 15 mmHg [23]. The detailed grading of the BHS standard is listed in Table 5.7.

	Percentage of errors		
Grade	<= 5 mmHg	<= 10 mmHg	<= 15 mmHg
A	60	85	95
B	50	75	90
C	40	65	85
D	Worse than grade C		

Table 5.7: Requirements for specific BP estimation device grades as given by the BHS standard [23].

Our system achieved grade B for SBP and grade A for DBP using the clinical dataset with maximum personalization, as shown by the percentages of errors in Figure 5.5.

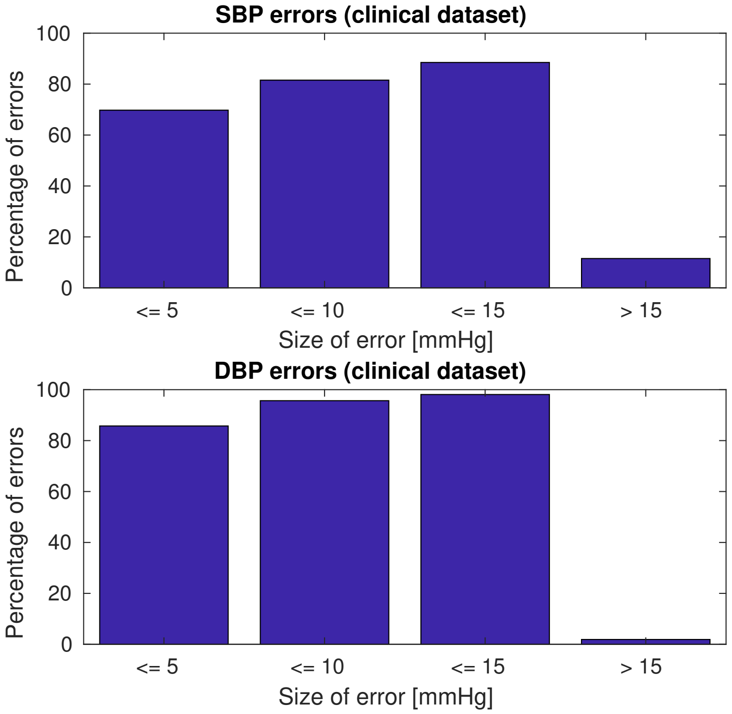


Figure 5.5: Percentages of errors under thresholds given by the BHS standard for the clinical dataset at maximum personalization.

With the more difficult everyday-life dataset, our system achieved grade C for SBP and grade A for DBP, as shown by the percentages of errors in Figure 5.6.

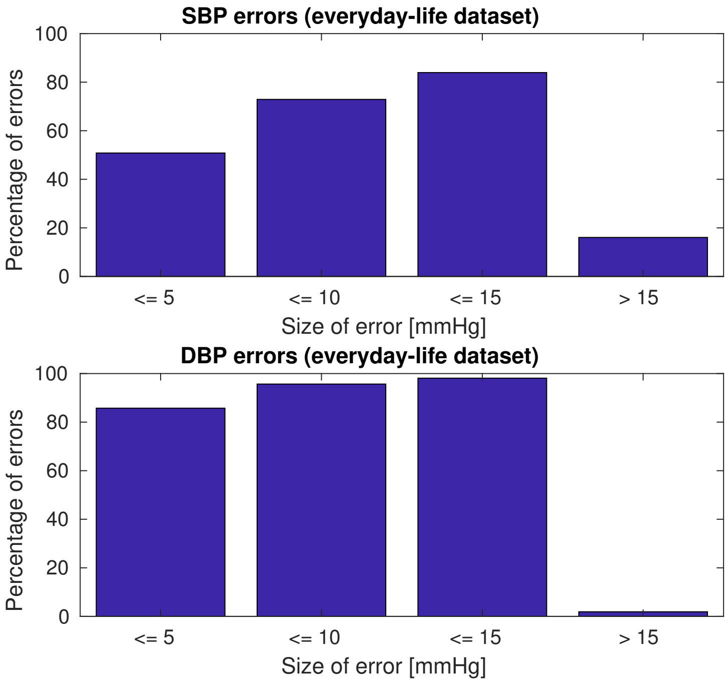


Figure 5.6: Percentages of errors under thresholds given by the BHS standard for the everyday-life dataset at maximum personalization.

5.4.1.2 Advancement of Medical Instrumentation (AAMI)

Second is the Advancement of Medical Instrumentation (AAMI) standard [23]. The AAMI requires BP measurement devices to have mean errors <= 5 mmHg and the corresponding standard deviations <= 8 mmHg, respectively. It was successfully met by our best-performing deep learning method at maximum personalization using the clinical dataset for both SBP and DBP. The everyday-life dataset lowest SBP errors were borderline acceptable, while DBP errors were fully acceptable.

Overall, deep learning regression model has shown the lowest MAE and is a prime candidate for use in production, followed by the Random Forest model.

A final observation is that the DBP errors and standard deviations are typically much lower compared to the SBP errors. This is expected, since DBP is generally more stable with less variations compared to SBP and is also in accordance with most related work discussed in Chapter 2.

Chapter 6

Web service

The architecture of the BP estimation system was briefly discussed in Chapter 3. To shortly summarize, the PPG signal is collected via a sensor embedded in a wristband and is then sent to a connected smartphone via bluetooth. Subsequently, the phone interacts with the predictive model through a web service, which then returns the BP prediction for the user, as illustrated in Figure 6.1. The reason behind the requirement of a web service lies in the high complexity of the signal preprocessing module, which depends on MATLAB and its toolboxes.

The current prototype uses a generic RESTful web service, which periodically receives a POST request with a JSON payload containing 30 seconds the raw PPG signal from the client. The signal is then saved into a .csv file. This format was chosen, as it is extremely generic and allows for interaction between software written in any programming language (e.g., MATLAB preprocessing module and Keras Python machine learning module, if deep learning model is used).

Once the raw PPG is saved by the web service, it is then read by the preprocessing module, which cleans the segment and obtains the high quality cycles. These are again saved into an intermediate .csv file.

Once the .csv with high quality cycles is created, the machine learning module is invoked. It computes the features and feeds them into the predic-

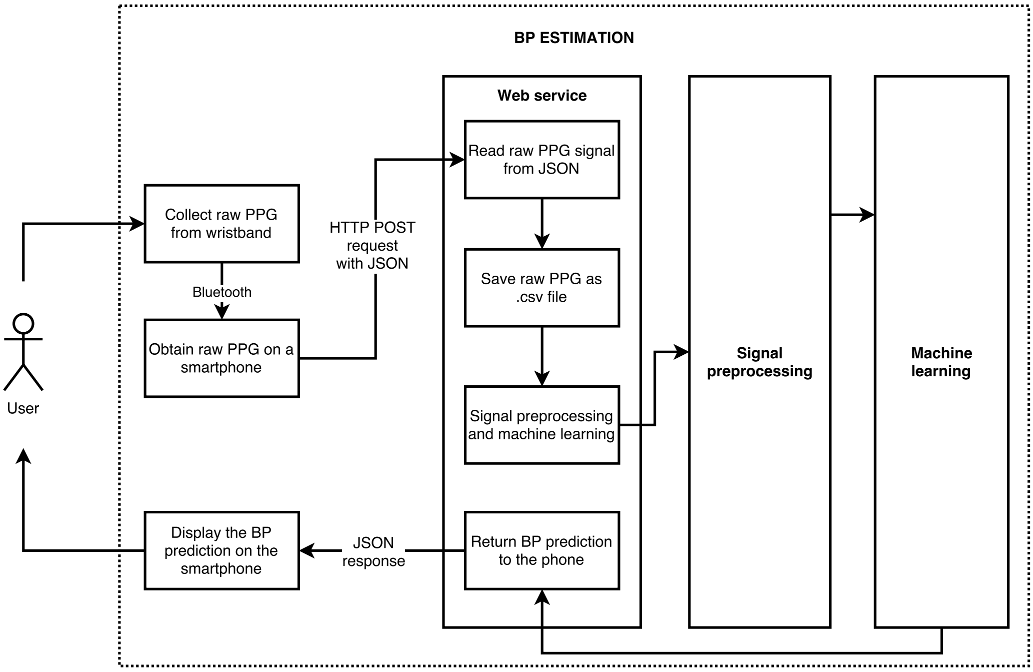

BP ESTIMATION
Web service
Read raw PPG signal from JSON
Collect raw PPG from wristband
HTTP POST request with JSON
Save raw PPG as .csv file
Bluetooth
Obtain raw PPG on a smartphone
Signal preprocessing
Machine learning
User
Signal preprocessing and machine learning
Display the BP prediction on the smartphone
JSON response
Return BP prediction to the phone

Figure 6.1: Schematic of the prototype system.

tive model. Currently, the machine learning module does a small amount of training of the general model, if ground-truth BP is provided as input alongside the PPG segment. This is done with the purpose of personalization of the model. Finally, the BP prediction is returned to the user's devices in the form of a JSON response.

Such a prototype design of the system allows for usage of any machine learning model independently of the preprocessing or wristband, making the system modular and robust. The whole process with a 30-second PPG segment without additional training takes around one second, making it viable for usage in real time.

Chapter 7

Conclusions

7.1 Summary

We have developed a continuous BP estimation system, which comprises two main modules, namely the signal preprocessing and machine learning module. The former is responsible for reading the PPG signal, detecting peaks and cycles, and cleaning the PPG signal of noisy segments, with the purpose of obtaining only the high-quality parts of the signal. The latter first calculates a number of features from the aforementioned high-quality PPG signal and feeds them into a machine learning model. This model does the regression and finally returns SBP and DBP predictions.

The system was validated on two datasets, one from a hospital environment (medical dataset from the MIMIC database) and the other from an everyday-life setting (collected at JSI). Several experiments were conducted, which evaluated the models' performance in terms of MAE. The most important was the LOSO experiment, which evaluated the generalization and robustness of the created models. It was determined that personalizing the models with some data from the specific subject greatly enhances the results.

Using the MIMIC clinical dataset, deep learning regression achieved the lowest $\mathrm{MAE}_{\mathrm{SBP}}$ of 5.61 mmHg and $\mathrm{MAE}_{\mathrm{DBP}}$ of 3.82 mmHg in the LOSO experiment. Both were achieved at maximum personalization, which corresponds to roughly 5-10 measurements (30 minutes) of subject-specific BP ground-truth alongside corresponding PPG signal used in the training.

With the everyday-life dataset, deep learning was not a viable algorithm because the dataset was too small. The lowest achieved MAE_{SBP} was 8.40 mmHg and the lowest MAE_{DBP} was 4.20 mmHg, this time using the Random Forest algorithm.

The two described models were complemented with a micro-batch RESTful web service for BP estimation, which allows for the interaction between the user's wristband and phone, and the model. The service periodically accepts a POST request with the raw PPG signal, which is then preprocessed by the first module. The result of the preprocessing is used by the second module to calculate the features, which are then fed into the regression model. The predictions returned by the model are forwarded back to the user's wristband.

7.2 Contributions

The main contribution of this thesis is the creation of a minimally obtrusive system for continuous BP estimation, using only the PPG signal. As the PPG sensor is embedded in a wristband, such a system allows for great freedom in the activities of the user and does not hinder or restrict the user in almost any way, while offering BP estimations on a near real-time basis. The main novelty lies in the merged signal preprocessing methodology, which combines state of the art approaches with custom modifications, creating a robust and effective preprocessing module capable of dealing with noisy data coming from most wristbands.

The secondary contribution is the successful validation of this approach using an elaborate experimental setting, which has shown, that the created models achieve low MAE, which mostly meet the requirements of the AAMI and BHS standard. The system was validated on two datasets, one from a clinical setting and another collected from 10 individuals during their everyday-life routine. So far, related work has focused on limited data, usually coming from databases or recorded in restricted controlled environments, using specialized equipment. This makes the collection and validation of the system on field collected data that much more valuable.

7.3 Discussion and limitations

One major limitation of our work arises from the low quality of the PPG signal collected with a wristband. It may happen that the preprocessing module removes nearly all the data due to its low quality. There is some possible regulation via hyper-parameters, which determine how strict the cleaning is, however, some signals can have extremely erratic movements within short windows, causing the preprocessing module to discard whole windows.

This in turn causes doubts about the performance of the system during physical activity, where the the contact between the PPG sensor of the wristband and the skin is continuously compromised, making the signal extremely noisy.

The second limitation comes from the fact, that we have not received a custom developed final version of the wristband with a display on time, as it was predicted in the project. This means that predictions cannot be shown directly on a wristband but rather on the phone, which is connected to the wristband.

The final limitation lies in the fact that deep learning regression has proven to be the best with clinical dataset, however, this algorithm was not validated using the everyday-life dataset due to its small size.

The developed prototype can in theory be used with any wristband capable of recording and saving PPG on a phone in a .csv file with the required format. This is possible since the web service is generic and the system is device independent. However, for the purpose of our experiments and evaluation, Empatica E4 wristband was used, so real life device independence was not yet validated other than with our experiments, which used the same system for two datasets originating from different devices.

In order for our system to meet the regulations on medical devices given by the European Commission, which would allow the device to be used in medical institutions in Europe, a large amount of additional time, effort, and financial support would be necessary, as the regulatory framework is extremely complex and requires a number of tests of the device, which go beyond just accuracy [58]. The regulatory framework requires exact classification of device, a study of user acceptance, a study of the level of invasiveness, etc. Furthermore, extensive and strict testing with a notable amount of real patients is mandatory, which in turn requires a participating medical institution and the approval of the ethics committee.

7.4 Future work

There are some possible continuations of our work, which will be additionally explored in the future.

First, more data should be collected from an everyday-life setting, allowing for deep learning methods to be applied on an everyday-life dataset.

Second, the methodology should be additionally tested and verified during physical activity, where the arm and wrist are moving, causing severe distortions in the signal.

Third, as we have created models for two datasets, which originate from the same problem domain and are thus related, transfer learning [59] could be considered. Transfer learning attempts to use the knowledge derived from one problem on another, when both problems are similar or originate from the same domain. Our two datasets are prime candidates for such an approach.

Finally, potential testing using hypertensive patients and their doctors should be conducted. The BP should be estimated using the developed system and be monitored with traditional cuff-based devices simultaneously. This would yield feedback from the doctors, which would be a useful indicator of whether such a system could be used by patients as a supplement or compliment to the traditional cuff-based devices.

Bibliography

[1] World Health Organization, The top 10 causes of death (Accessed September 4th, 2017).
URL `http://www.who.int/mediacentre/factsheets/fs310/en/`

[2] E. M. Frese, H. S. Sadowsky, Blood pressure measurement guidelines for physical therapists, Cardiopulmanory Physical Therapy Journal 22 (2) (2011) 5–12.

[3] X. He, R. A. Goubran, X. P. Liu, Evaluation of the correlation between blood pressure and pulse transit time, in: 2013 IEEE International Symposium on Medical Measurements and Applications (MeMeA), 2013, pp. 17–20.

[4] U. Maeda, B.-J. Shen, E. R. Schwarz, K. A. Farrell, S. Mallon, Self-efficacy mediates the associations of social support and depression with treatment adherence in heart failure patients, International Journal of Behavioral Medicine 20 (1) (2013) 88–96.

[5] J. Lee, K. Matsumura, K. i. Yamakoshi, P. Rolfe, S. Tanaka, T. Yamakoshi, Comparison between red, green and blue light reflection photoplethysmography for heart rate monitoring during motion, in: 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2013, pp. 1724–1727.

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

[7] J. Allen, Photoplethysmography and its application in clinical physiological measurement, Physiological Measurement 28 (3) (2007) R1.

[8] L. A. Geddes, M. H. Voelz, C. F. Babbs, J. D. Bourland, W. A. Tacker, Pulse transit time as an indicator of arterial blood pressure, psychophysiology 18 (1) (1981) 71–74.

[9] R. P. Smith, J. Argod, J.-L. Pépin, P. A. Lévy, Pulse transit time: an appraisal of potential clinical applications, Thorax 54 (5) (1999) 452–457.

[10] C. C. Y. Poon, Y. T. Zhang, Cuff-less and noninvasive measurements of arterial blood pressure by pulse transit time, in: 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference, 2005, pp. 5877–5880.

[11] M. Kachuee, M. M. Kiani, H. Mohammadzade, M. Shabany, Cuffless blood pressure estimation algorithms for continuous health-care monitoring, IEEE Transactions on Biomedical Engineering 64 (4) (2017) 859–869.

[12] X. F. Teng, Y. T. Zhang, Continuous and noninvasive estimation of arterial blood pressure using a photoplethysmographic approach, in: Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (IEEE Cat. No.03CH37439), Vol. 4, 2003, pp. 3153–3156 Vol.4.

[13] Y. Kurylyak, F. Lamonaca, D. Grimaldi, A neural network-based method for continuous blood pressure estimation from a ppg signal, in: 2013 IEEE International Instrumentation and Measurement Technology Conference (I2MTC), 2013, pp. 280–283.

[14] F. Lamonaca, K. Barbe, Y. Kurylyak, D. Grimaldi, W. V. Moer, A. Furfaro, V. Spagnuolo, Application of the artificial neural network for blood pressure evaluation with smartphones, in: 2013 IEEE 7th International Conference on Intelligent Data Acquisition and Advanced Computing Systems (IDAACS), Vol. 01, 2013, pp. 408–412.

[15] R. Banerjee, A. Ghose, A. D. Choudhury, A. Sinha, A. Pal, Noise cleaning and gaussian modeling of smart phone photoplethysmogram to improve blood pressure estimation, in: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015, pp. 967–971.

[16] X. Xing, M. Sun, Optical blood pressure estimation with photoplethysmography and fft-based neural networks, Biomedical Optics Express 7 (8) (2016) 3007–3020.

[17] J. Lázaro, E. Gil, J. M. Vergara, P. Laguna, Pulse rate variability analysis for discrimination of sleep-apnea-related decreases in the amplitude fluctuations of pulse photoplethysmographic signal in children, IEEE Journal of Biomedical and Health Informatics 18 (1).

[18] Q. Li, G. D. Clifford, Dynamic time warping and machine learning for signal quality assessment of pulsatile signals, Physiological Measurement 33 (9).

[19] M. Robnik-Šikonja, I. Kononenko, Theoretical and empirical analysis of ReliefF and RReliefF, Machine Learning 53 (1-2).

[20] L. Billard, E. Diday, Symbolic Regression Analysis, Springer Berlin Heidelberg, 2002.

[21] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, 2016, http://www.deeplearningbook.org.

[22] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, H. E. Stanley, PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals, Circulation 101 (23) (2000) e215–e220.

[23] E. O'Brien, N. Atkins, Validation and Reliability of Blood Pressure Monitors, Humana Press, Totowa, NJ, 2007.

[24] J. Handler, The importance of accurate blood pressure measurement, The Permanente Journal (2009) 51–54.

[25] J. R. Sowers, M. Epstein, E. D. Frohlich, Diabetes, hypertension, and cardiovascular disease, Hypertension 37 (4) (2001) 1053–1059.

[26] Accurate blood pressure measurement for medical instrumentation (Accessed August 31st, 2017).
URL `http://accuratebloodpressure.com/blood_pressure.html`

[27] M. Kay, J. Santos, M. Takane, mhealth: New horizons for health through mobile technologies, World Health Organization 64 (7) (2011) 66–71.

[28] T. Tamura, Y. Maeda, M. Sekine, M. Yoshida, Wearable photoplethysmographic sensors—past and present, Electronics 3 (2) (2014) 282–302.

[29] Empatica Inc., Real-time physiological signals - e4 eda/gsr sensor. (Accessed September 1st, 2017).
URL `https://www.empatica.com/e4-wristband`

[30] C. Sammut, G. I. Webb, Encyclopedia of Machine Learning, Springer US, 2010.

[31] Wikipedia contributors, Apple watch — wikipedia, the free encyclopedia (Accessed January 14th, 2018).
URL `https://en.wikipedia.org/w/index.php?title=Apple_Watch&oldid=820429150`

[32] Wikipedia contributors, Microsoft band 2 — wikipedia, the free encyclopedia (Accessed January 14th, 2018).
URL https://en.wikipedia.org/w/index.php?title=Microsoft_Band_2&oldid=814706092

[33] K. H. Shelley, Photoplethysmography: Beyond the calculation of arterial oxygen saturation and heart rate, Anesthesia Analgesia 105 (6).

[34] M. Elgendi, On the analysis of fingertip photoplethysmogram signals, Current Cardiology Reviews 8 (1).

[35] G. E. P. Box, G. Jenkins, Time Series Analysis, Forecasting and Control, Holden-Day, Incorporated, 1990.

[36] Wikipedia contributors, Heart rate — wikipedia, the free encyclopedia (Accessed January 14th, 2018).
URL https://en.wikipedia.org/w/index.php?title=Heart_rate&oldid=819816330

[37] T. W. Parks, C. S. Burrus, Digital Filter Design, Wiley-Interscience, New York, NY, USA, 1987.

[38] The MathWorks, Inc., Correlation coefficients (Accessed December 5th, 2017).
URL https://www.mathworks.com/help/matlab/ref/corrcoef.html

[39] D. J. Berndt, J. Clifford, Using dynamic time warping to find patterns in time series, in: Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining, 1994, pp. 359–370.

[40] C. M. Bishop, Pattern Recognition and Machine Learning (Information Science and Statistics), Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[41] K. Najarian, R. Splinter, Biomedical Signal and Image Processing, Second Edition, CRC Press, Taylor & Francis Group, 2012.

[42] M. Robnik-Šikonja, I. Kononenko, Theoretical and empirical analysis of relieff and rrelieff, Machine Learning 53 (1) (2003) 23–69.

[43] N. R. Draper, H. Smith, Applied Regression Analysis, Third Edition, 3rd Edition, Wiley Series in Probability and Statistics, Wiley-Interscience, 1998.

[44] L. Breiman, J. Friedman, R. A. Olshen, C. J. Stone, Classification and regression trees, The Wadsworth statistics / probability series, CRC, 1984.

[45] T. Hastie, R. Tibshirani, J. Friedman, The elements of statistical learning: data mining, inference and prediction, 2nd Edition, Springer, 2009. URL `https://web.stanford.edu/~hastie/Papers/ESLII.pdf`

[46] J. R. Koza, Genetic programming as a means for programming computers by natural selection, Statistics and Computing 4 (2).

[47] A. E. Eiben, J. E. Smith, Introduction to Evolutionary Computing, SpringerVerlag, 2003.

[48] G. Sannino, I. D. Falco, G. D. Pietro, Genetic programming for a wearable approach to estimate blood pressure embedded in a mobile-based health system, in: 2015 IEEE 27th International Conference on Tools with Artificial Intelligence (ICTAI), 2015.

[49] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, D. Hassabis, Mastering the game of go with deep neural networks and tree search, Nature 529 (2016) 484–503.

URL http://www.nature.com/nature/journal/v529/n7587/full/nature16961.html

[50] M. A. Nielsen, Neural Networks and Deep Learning, Determination Press, 2015.

[51] MATLAB, version 9.2.0 (R2017a), The MathWorks Inc., Natick, Massachusetts, 2017.

[52] F. Chollet, et al., Keras, https://github.com/keras-team/keras (2015).

[53] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, X. Zheng, TensorFlow: Large-scale machine learning on heterogeneous systems, software available from tensorflow.org (2015).
URL https://www.tensorflow.org/

[54] A. Ng, Neural networks and deep learning (Accessed December 27th, 2017).
URL https://www.coursera.org/learn/neural-networks-deep-learning

[55] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, CoRR abs/1412.6980.

[56] E. O'Brien, B. Waeber, G. Parati, J. Staessen, M. G. Myers, Blood pressure measuring devices: recommendations of the European Society of Hypertension, British Medical Journal 322 (7285).

[57] Omron Healthcare Inc., Clinical validation (Accessed January 28th, 2018).
URL `https://omronhealthcare.com/service-and-support/clinical-validation/`

[58] European Commission, Regulatory framework for medical devices (Accessed January 28th, 2018).
URL `https://ec.europa.eu/growth/sectors/medical-devices/regulatory-framework_sl`

[59] S. J. Pan, Q. Yang, A survey on transfer learning, IEEE Transactions on Knowledge and Data Engineering 22 (10).