Continuous Blood Pressure Estimation from PPG Signal

Gašper Slapničar and Mitja Luštrek
Joef Stefan Institute, Jamova cesta 39, 1000 Ljubljana
E-mail: gasper.slapnicar@ijs.si, mitja.lustrek@ijs.si

Matej Marinko
Faculty of Mathematics and Physics, Jadranska cesta 19, 1000 Ljubljana
E-mail: matejmarinko123@gmail.com

Keywords: photoplethysmography, blood pressure estimation, regression analysis, m-health

Received: November 11, 2017

Given the importance of blood pressure (BP) as a direct indicator of hypertension, regular monitoring is encouraged for healthy people and mandatory for patients at risk from cardiovascular diseases. We propose a system in which photoplethysmogram (PPG) is used to continuously estimate BP. A PPG sensor can be easily embedded in a modern wearable device, which can be used in such an approach. The PPG signal is first preprocessed in order to remove major noise and movement artefacts present in the signal. A set of features describing the PPG signal on a per-cycle basis is then computed to be used in regression models. The predictive performance of the models is improved by first using the RReliefF algorithm to select a subset of relevant features. Afterwards, personalization of the models is considered to further improve the performance. The approach is validated using two distinct datasets, one from a hospital environment and the other collected during every-day activities. Using the MIMIC hospital dataset, the best achieved mean absolute errors (MAE) in a leave-one-subject-out (LOSO) experiment were 4.47 mmHg for systolic and 2.02 mmHg for diastolic BP, at maximum personalization. For everyday-life dataset, the lowest errors in the same LOSO experiment were 8.57 mmHg for systolic and 4.42 mmHg for diastolic BP, again using maximum personalization. The best performing algorithm was an ensemble of regression trees.

Povzetek: Krvni tlak je neposreden pokazatelj hipertenzije. Razvili smo sistem, ki krvni tlak ocenjuje iz fotopletizmograma (PPG), kakršen je že vgrajen v večino modernih senzorskih zapestnic. Signal PPG smo sprva predprocesirali in segmentirali na cikle. Predprocesiranje odpravi večino šuma, ki se pogosto pojavlja zaradi gibanja. Iz očiščenega signala smo nato izračunali množico značilk, ki smo jih uporabili v regresijskih modelih. Sistem smo izboljšali z uporabo algoritma RReliefF za izbor relevantnih 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 smo model vsakič naučili na vseh osebah razen eni in ga nato testirali na izpuščeni osebi. Z uporabo klinične podatkovne množice smo v omenjenem poizkusu dosegli najnižji povprečni absolutni napaki (MAE) 4.47 mmHg za sistolični in 2.02 mmHg za diastolični krvni tlak, pri največji stopnji personalizacije. Za množico, zbrano med dnevnimi aktivnostmi, smo dosegli najnižji napaki 8.57 mmHg za sistolični in 4.42 mmHg za diastolični krvni tlak, ponovno pri največji stopnji personalizacije. Najbolje se je obnesel ansambel regresijskih dreves.

1 Introduction

World Health Organization (WHO) listed cardiovascular diseases as the most common cause of death in 2015, responsible for almost 15 million deaths combined [1]. Hypertension is one of the most common precursors of such diseases and can be easily detected with regular blood pressure (BP) monitoring, which is especially critical for patients already suffering from hypertension or related cardiovascular diseases, as it can indicate potential vital threats to their health.

While regular BP monitoring is important, it is also troublesome, as devices using inflatable cuffs are still considered the "golden standard". The cuff placement is critical, as the sensor must be located directly above the main artery in the upper arm area, at approximately heart height [4]. These requirements impose relatively strict movement restrictions on the subject and require substantial time commitment, thus causing low subject adherence to regular monitoring. Furthermore, when done by the subject him/herself in a home environment, this process can cause stress, which in turn influences the BP values, making the measurements less reliable. This problem is usually not alleviated by having the medical personnel perform the measurement, as this can again cause anxiety in the subject, commonly known as the "white coat syndrome".

Our work focuses on photoplethysmogram (PPG) analysis and the development of a robust non-obtrusive method for continuous BP estimation. It will be implemented and used in an m-health system based on a wristband with an embedded PPG sensor. This will allow the user to wear the device without any interference or limits imposed upon their daily routine, allowing for truly continuous measuring without stressing the user and thus potentially influencing the BP values.

The rest of the paper is organized as follows. Section 2 gives a brief overview of the related work. Section 3 explains the methodology we have used, focusing on signal pre-processing and machine learning features. Section 4 elaborates on the experimental setup and results, and Section 5 concludes with a summary and plans for future work.

2 Related work

Photoplethysmography is a relatively simple technique based on inexpensive technology, which is becoming increasingly popular in wearable devices for heart rate estimation. It is based on the illumination of the skin and measurement of changes in its light absorption [5]. In its basic form it only requires a light source to illuminate the skin (typically a light-emitting diode – LED light) and a photodetector (photodiode) to measure the amount of light either transmitted through, or reflected from the skin. Thus PPG can be measured in either transmission or reflectance mode. Both modes of operation are shown in Figure 1.

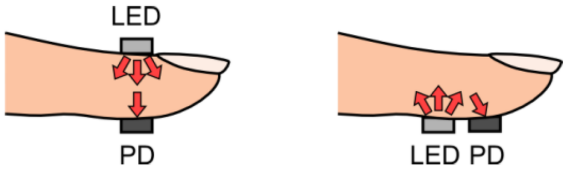

Figure 1: Transmission and reflectance mode in which the PPG signal can be obtained. LED is the light source while PD is the photodetector [6].

With each cardiac cycle, the heart pumps blood towards the periphery of the body. This produces a periodic change in the amount of light that is either absorbed or reflected from the skin to the photodetector, as the tissue changes its tone based on the amount of blood in it.

Exploring the recent applications of PPG, we can see that it is becoming more widely used in BP estimation. One of two common approaches are typically used:

1. BP estimation using two sensors (PPG + Electrocardiogram (ECG))
2. BP estimation using the PPG sensor only

The first approach requires the use of two sensors, typically an ECG and a PPG sensor, in order to measure the time it takes for a single heart pulse to travel from the heart to a peripheral point in the body. This time is commonly known as pulse transit time (PTT) or pulse arrival time (PAT), and its correlation with BP changes is well established.

The more recent approach is focused on the PPG signal only; however, the relationship between the PPG and BP is only postulated and not as well established as the relationship between the PTT and BP. This approach is, however, notably less obtrusive, especially since PPG sensors have recently become very common in most modern wristbands.

BP is commonly measured in millimeters of mercury (mmHg), which is a manometric unit routinely used in medicine and many other scientific fields. A mercury manometer is a curved tube containing mercury, which is closed at one end while pressure is applied on the other end. 1 mmHg of pressure means that the pressure is large enough to increase the height of the mercury in the tube for 1 mm. To put the values discussed in this paper into perspective, the normal healthy adult BP is considered to be around 120 mmHg (16 kPa) for systolic and 80 mmHg (11 kPa) for diastolic BP [2].

One of the earliest PPG-only attempts was conducted by Teng et. al. in 2003 [3]. The relationship between the arterial BP (ABP) and certain features of the PPG signals was analyzed. Data were obtained from 15 young healthy subjects in a highly controlled laboratory environment, ensuring constant temperature, no movement and silence. The mean differences between the linear regression estimations and the measured BP were 0.21 mmHg for systolic (SBP) and 0.02 mmHg for diastolic BP (DBP). The corresponding standard deviations were 7.32 mmHg for SBP and 4.39 mmHg. Using mean errors instead of mean absolute errors as the evaluation metric is questionable, since it does not reflect the actual performance of the derived model and the error can be extremely low, even if the actual predictions are high above and under the actual observed BP values.

A paper was published in 2013 in which the authors used data from the Multiparameter Intelligent Monitoring in Intensive Care (MIMIC) waveform database [7, 8] to extract 21 time domain features and use them as an input vector for artificial neural networks (ANNs) [9]. The results are not quite as good as with the linear regression model described earlier; however, the data was obtained from a higher number and variety of patients in a less controlled environment. Mean absolute errors of less than 5 mmHg for both SBP and DBP were reported. While the environment was less controlled compared to the previous work, the patients were still within a hospital setting and hospital equipment was used for data collection. Furthermore, only an undisclosed subset of all the available data from MIMIC was used.

Another research was conducted in 2013 in which the authors used a smartphone camera to capture the PPG signal using the camera flash as the light source and the phone camera as the photodiode [10]. PPG features were again extracted and fed to a neural network, which estimated 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. It is not clear how many subjects participated in the experiment, however, they reported the maximum error not exceeding 12 mmHg. The error metric is not explained in detail, however, based on the given results table, we can presume that MAE was used. Such a method requires some 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.

It is clear that the PPG-only approach has potential, however, a robust unobtrusive method that works well on a general case is yet to be developed.

3 Methodology

The proposed system consists of two main modules, namely the signal pre-processing and machine learning module. The former is responsible for cleaning the PPG signal of most noise and then segmenting it into cycles, where one PPG cycle corresponds to a single heart beat. The latter extracts features describing the PPG signal on a per-cycle basis, selects a subset of relevant features using the RReliefF algorithm [12], and finally feeds the subset into regression algorithms, which build the prediction models.

3.1 Signal pre-processing

PPG sensors must be very sensitive in order to detect tiny variations in light absorption of the tissue. This also makes them highly susceptible to movement artefacts. This problem is especially obvious when dealing with PPG collected via a wristband, as the contact between the sensor and the skin can be compromised during arm movements. This is partially alleviated by using green light, which is less prone to artefacts, however, major artefacts often remain in the signal. Subsequently, substantial effort is directed towards PPG pre-processing.

3.1.1 Cleaning based on established medical criteria

In the first phase, both the BP and PPG signal are roughly cleaned based on established medical criteria [13]. A 5-second sliding window is used to detect segments with extreme BP values or extreme changes of the BP in a short time period. Thresholds for extreme values and changes are selected based on established medical criteria in related work [13] and are given in Table 1. Some thresholds were slightly modified, since the criteria given in the referenced paper seem too strict for some subjects encountered in our datasets. We have thus loosened the criteria in accordance with empirical observations in our datasets (e.g., the original criteria excludes all data with SBP > 180, while we observed some segments with SBP over 180 mmHg).

After the cleaning of the clinical dataset, 85% of data is kept on average, while 15% is discarded. This is very subject dependent, as for some subjects nearly all the data is removed (e.g., sensor anomaly which shows 0 ABP almost all the time), while for majority of subject most of the data is kept. For everyday-life dataset, which contains a lot more noise, only 40% of data is kept, while 60% is discarded. This is the result of some subjects having long noisy segments of the PPG signal. It should be noted, that these percentages are also subject of the parameters for trade-off between the required quality and the amount of signal kept, which are discussed in 3.1.3.

Criterion	Threshold
SBP	> 250 or < 80
DBP	> 150 or < 40
SBP – DBP	< 20
ΔSBP or ΔDBP in 5 sec	> 50

Table 1: Established medical criteria and thresholds for rough signal cleaning. Δ signifies a change in BP value. 5-second segments meeting any of these criteria are removed from the signal.

3.1.2 Peak and cycle detection

In order to do further cleaning and subsequent feature extraction, PPG cycle detection is mandatory. This is not trivial, as substantial noise in the PPG signal poses a significant problem, as mentioned earlier.

This problem was tackled in several steps. First, a filtering transformation, which enhances the systolic upslopes of the pulses in the PPG signal, is used. It is designed to use the derivative of the PPG signal at lower frequencies, in order to detect the abrupt upslopes of the systolic pulse compared to the diastolic or dicrotic pulse in the PPG signal. This is based on a low-pass differentiator (LPD) filter, which removes high frequency components and performs differentiation. Once the steepest points in the PPG signal are located, the following peak is chosen as the PPG systolic peak. Afterwards, a time-varying threshold for peak detection is applied, which ensures that potential double peaks or diastolic peaks close to the systolic ones are not chosen. The procedure is explained in detail in a paper by Lzaro et al. [14].

After the peaks are detected, finding the cycle start-end points is simpler, as the dominant valleys between the detected peaks must be found. An example of detected peaks and cycle locations using the described method is shown in Figure 2.

3.1.3 Cleaning based on ideal templates

After cycles are successfully detected, the second cleaning phase begins. A 30-second sliding window is used.

First, the most likely length of a cycle L in the current window is determined using autocorrelation analysis. A copy of the PPG signal in the current window is taken and

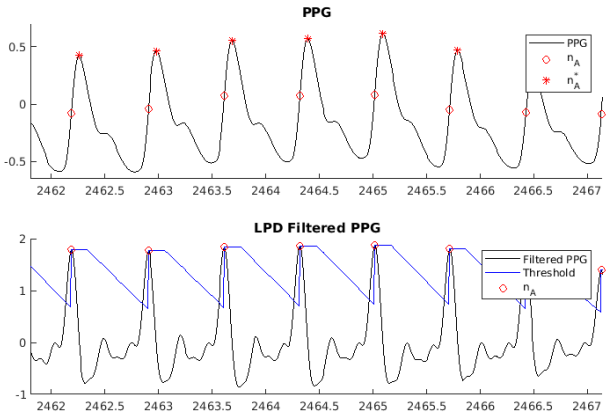

Figure 2: The upper subplot shows a PPG segment. Lower subplot shows the LPD filtering transformation of the same PPG segment. Peaks of the transformation in the lower subplot correspond to the steepest systolic upslopes of the PPG in the upper subplot, and are denoted as n_A. Actual detected PPG peaks are denoted as n_A^*.

shifted sample by sample up to a certain length that contains at least two heart beats. When the copy is shifted by the number of samples corresponding to exactly one cycle, the autocorrelation reaches its first peak, and this number of samples is chosen as L.

Presuming that the majority of cycles within a 30-second window are not morphologically altered, we can create an "ideal cycle template" for this window. Such a template is created by always taking the next L samples from each cycle starting point and then computing the mean cycle. Each individual cycle is then compared to the computed template and its quality is evaluated using three signal quality indices (SQIs), which are defined as follows [15]:

1. **SQI1**: First L samples of each cycle are taken and then each cycle is directly compared to the template using a correlation coefficient.

2. **SQI2**: Each cycle is interpolated to length L and then the correlation coefficient with the template is computed.

3. **SQI3**: The distance between each cycle and the template is computed using dynamic time warping (DTW).

Finally the thresholds for each SQI are empirically determined. Each cycles' SQIs are evaluated and if they reach the required quality threshold, that cycle is kept, otherwise it is removed. If more than half the cycles in the current 30-second window are under the thresholds, the whole window is discarded as too noisy. An example of this cleaning is shown in Figure 3.

Once the PPG signal is cleaned and only high-quality cycles with minimal morphological anomalies remain, features can be extracted from each cycle.

3.2 Machine learning

In order to derive the relationship between the PPG and BP, features describing the PPG signal were computed and then the relevant subset of these features was selected to be used in the regression algorithms.

3.2.1 Features

In accordance with the related work [3, 9, 10], several time-domain features were computed from the PPG signal, and the set of features was further expanded with some from the frequency [13] and complexity-analysis domains. Most features focus on describing the morphology of a given PPG cycle, as shown in Figure 4.

Feature	Description
Tc	Cycle duration
Ts	Time from start of cycle to systolic peak
Td	Time from systolic peak to end of cycle
Tnt	Time from systolic peak to diastolic rise
Ttn	Time from diastolic rise to end of cycle
S1	Area under the curve (AUC) from start of cycle to max upslope point
S2	AUC from max upslope point to systolic peak
S3	AUC from systolic peak to diastolic rise
S4	AUC from diastolic rise to end of cycle
AUC syst	S1 + S2
AAC syst	Area above the curve (AAC) from start of cycle to systolic peak
AUC diast	S3 + S4
AAC diast	AAC from systolic peak to end of cycle

Table 2: Elaborations of some of the used features shown in Figure 4.

In addition to the features focusing on the PPG cycle morphology, which were highlighted thus far, the following features were computed and considered:

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

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

2. *LASI – Large Artery Stiffness Index*: an indicator of arterial stiffness, which is denoted as *Tnt* in Figure 4 and Table 2.

3. *Complexity analysis*: signal complexity and mobility are computed for the 30-second PPG segment containing the current cycle. Mobility represents an estimate of the mean frequency and is proportional to the standard deviation of the power spectrum. Complexity gives an estimate of change in frequency by comparing the signal similarity to a pure sine wave. They are

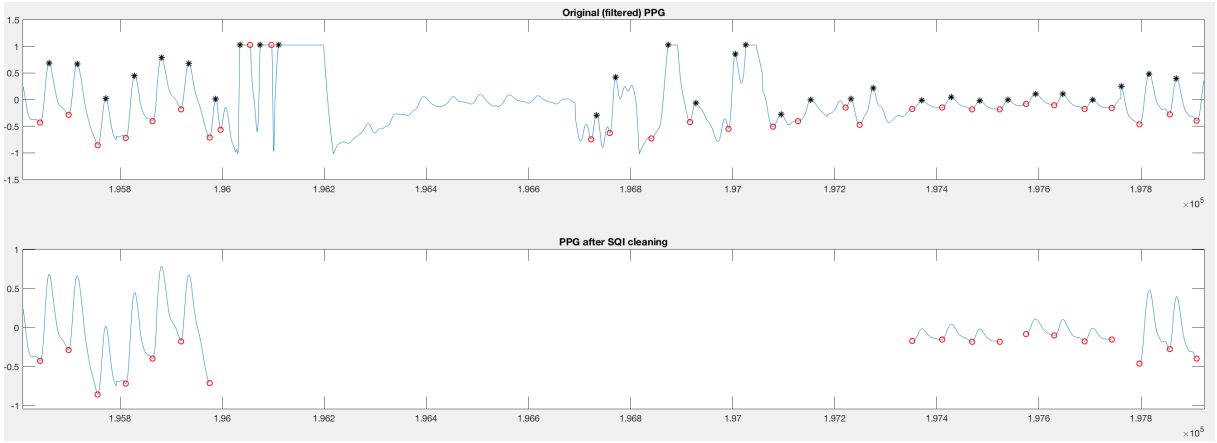

Figure 3: Example of the cleaning algorithm in the second phase of the signal pre-processing. Comparing the top (uncleaned) and bottom (cleaned) PPG signal, we see that the obvious artefact segments are removed.

given by Najarian and Splinter [11] as follows (presuming a zero-mean signal):

$$S_0 = \sqrt{\frac{\sum_{i=1}^{N} x_i^2}{N}},$$

$$S_1 = \sqrt{\frac{\sum_{j=2}^{N-1} d_j^2}{N-1}},$$

$$S_2 = \sqrt{\frac{\sum_{k=3}^{N-2} g_k^2}{N-2}},$$

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

$$Mobility = \sqrt{\frac{var(d)}{var(x)}} = \frac{S_1}{S_0}$$

$$Complexity = \frac{mobility(d)}{mobility(x)} = \sqrt{\frac{S_2^2}{S_1^2} - \frac{S_1^2}{S_0^2}},$$

4. *FFT features*: amplitudes and phases of the frequency-domain representation of the 30 second PPG segment containing the current cycle. The length of the window was chosen such that it contains enough cycles (expected 1 cycle per second) for the frequencies in the segment to be reliably determined.

Considering all the time and frequency-domain features along with the complexity-analysis features, and the amount of instances (cycles) available, we are often dealing with a very large matrix of training data. The number of rows (instances) is on the order of magnitude 10^5 and the number of columns (features) is on the order of magnitude 10^2, thus dimensionality reduction through selection of a subset of relevant features is feasible, but not mandatory. More importantly, feature selection allows us to determine which features are useful for the learning process, and which are irrelevant, allowing us to obtain a smaller subset containing only the relevant features.

3.2.2 Feature selection

The RReliefF algorithm was chosen for feature selection. It is a modification of the ReliefF algorithm, suitable for regression problems with continuous target variables. The algorithm was applied to a subset of 10% of all data chosen randomly. This was repeated 10 times. All the features with non-zero relevance, as chosen by the algorithm, were considered in each iteration and their importance was saved. Looking at the final scores of the algorithm across all the iterations, we notice that quite a few features are considered irrelevant, while the same features are commonly chosen as important for both SBP and DBP, as shown in Figure 5. Noting the fact that the same features were selected in each of the 10 iterations, we can assume that the relevant features are not dependent on the selected subset of the available data.

As mentioned, all the features with non-zero importance were taken, as more than half were discarded as irrelevant by the RReliefF algorithm. Among the non-zero importance features, some features from each of the groups mentioned earlier (temporal, frequency and complexity analysis) are present. Most area-based features were marked as irrelevant, while certain times (Tc, Ts and Td), both complexity-analysis (signal complexity and signal mobility) as well as some frequency-domain (amplitudes and phases at low frequencies) features were marked as important. These non-zero importance features were then used in the regression algorithms.

The relevant features were determined using the larger and more varied dataset from the MIMIC database. The same subset of features was also used with the smaller everyday-life dataset. Both datasets are described in more detail in the following section.

Since the feature selection procedure only slightly improved the results, we have not considered experiments with other or additional feature selection methods.

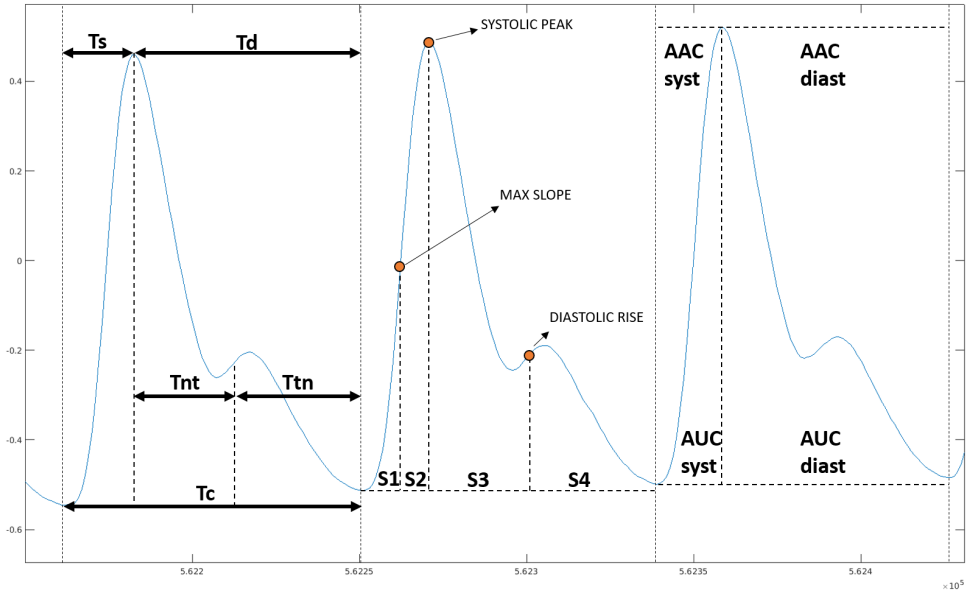

Figure 4: Time and area based features that describe the morphology of the PPG signal on a per-cycle basis. The features are listed and elaborated in Table 2.

4 Experiments and results

In an effort to make the proposed method as general as possible, two datasets were considered for the experimental evaluation. The data from all subjects, which met the requirements of having both the PPG and BP signals recorded, were always used in the experiments.

4.1 Data

The first dataset is from the publicly accessible MIMIC database, which is commonly used for experiments and competitions in the signal processing field. The original version contains data from 72 hospitalized patients. All patients with both the PPG and BP signal were originally considered, however, after the filtering and pre-processing, only 41 patients had enough high-quality data remaining to be used in the experiments. All the data was collected in a hospital environment using hospital measuring equipment, including an ABP measuring device. The ABP is measured by inserting a cathether in an artery, making it highly invasive, however, it offers the most precise BP monitoring.

The second dataset was collected at Jožef Stefan Institute (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. This device is reported to be clinically validated according to the British Hypertension Society and the Association for the Advancement of Medical Instrumentation (AAMI) protocols [17], which means that the mean errors do not exceede 5 mmHg. The collection procedure at JSI was conducted in accordance with the standardized clinical protocol. The correct placement of the cuff on the upper arm area with the sensor above the main artery was ensured. The measurements were done in an upright sitting position, making sure the cuff was located at approximately heart height. The recommended protocol was followed as best as possible, however, in an ideal situation the ground truth BP should be measured as ABP within an artery. Due to the invasive nature of 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.

In the first completed phase of the data collection, 8 healthy subjects were considered, 5 male and 3 female. Each subject wore the wristband PPG measuring device for several hours during their everyday activities. They measured their BP every 30 minutes or more often. Finally, 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. Furthermore, additional physiological variations (e.g., breathing rate) could be obtained from the PPG and used for the BP estimation, however, this was not yet considered but might be a subject of future work.

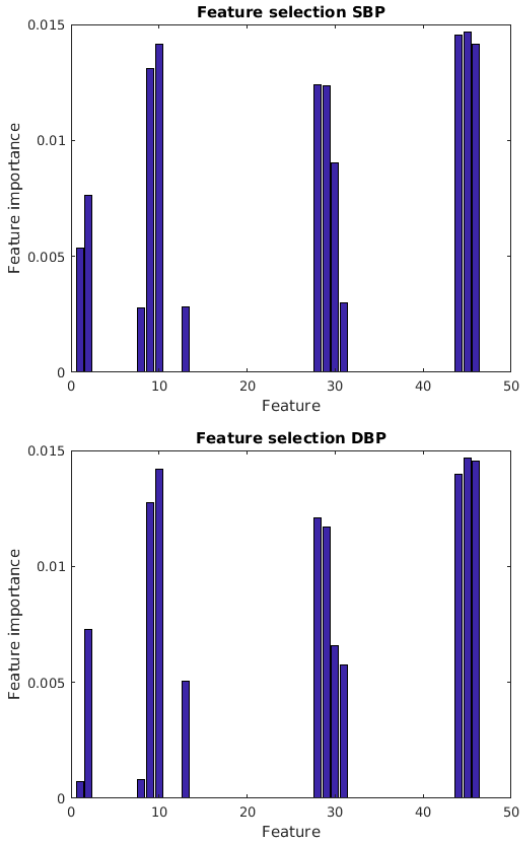

Figure 5: The output of RReliefF algorithm, which shows the feature importance for each of the considered features.

4.2 Experimental setup

Two experimental setups were considered, 5-fold cross validation and LOSO. The purpose of the first was to establish initial observations about the selected features and performance of different regression algorithms. The second experiment was conducted to evaluate the generalization performance of the algorithms and subsequently determine potential requirement for personalization.

4.2.1 K-fold cross validation

The MIMIC dataset consisted of roughly 200 000 instances post filtering, which correspond to 41 patients. The instances were obtained by uniformly taking 20 3-minute segments from the whole recording for a given patient. Each instance (cycle) in a given 3-minute segment was assigned the mean SBP and mean DBP of this segment. This simulates the patients measuring their BP periodically, but not more than once in 3 minutes.

K-fold cross validation ($k = 5$) was conducted, where instances were first shuffled randomly and then all the data was split into nearly equal folds. Then $k - 1$ folds were taken for learning and the remaining fold was used for testing. This was repeated k times. The random shuffling of instances makes it so that instances belonging to a given subject might appear in both training and testing sets. This was taken into account (a sort of implicit personalization), as this experiment was merely a starting point to determine the initial performance of the algorithms and was later complemented by a Leave-one-subject-out experiment.

Several regression algorithms were compared using the full set of features. The algorithm that performed best using all the features was additionally evaluated using only the subset of best features as selected by the RReliefF algorithm. The predictive performance of these options in 5-fold cross validation is discussed in detail in the Results section.

4.2.2 Leave-one-subject-out

Due to increased computational complexity of a leave-one-subject-out experiment compared to k-fold cross-validation, data was additionally sub-sampled, by taking 500 uniformly selected cycles from each patient's data.

During the initial attempt, a regression model was trained in each iteration on all the subjects, except the one left out. It was ensured that no instances from the testing subject appeared in the training set. This yielded poor results. Notable improvements can be made by using a small amount of each patient's data for training, most likely due to each patient having a subtly unique cardiovascular dynamic and relation between PPG and BP. This 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 models.

In the second attempt, the regression models were again trained using all the subjects except the one left out. This time, however, the models were further personalized by 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 instances was taken from the testing data and used in training in order to personalize the model to the current patient. This ensures personalization with different BP values, as taking just a single group of instances gives little information, since the BP will be constant within this group. Given the fact that the MIMIC data consists of roughly 5x the number of patients compared to everyday-life data, the personalization data for it was multiplied 5 times, making it noticeable within the large amount of training data from the remaining patients.

During both attempts, several regression algorithms were once again considered, as given in Tables 3 and 4. The MAE was used as the evaluation metric. All models were compared with a dummy regressor, which always predicted the mean BP value of the same combination of general and personalization data as the other models used for training. Finally, the regressor with the lowest MAE was chosen as

best.

For successful personalization, the user should measure their PPG continuously and also make a few periodic measurements of their BP using a reliable commercial device. This allows the model to personalize to the user, learning from a small sample of their labeled data, thus improving its predictive performance.

4.3 Results

Using the personalization approach, notable improvements have been made over the dummy regressor in both experiments. The results are discussed in detail in the following sections.

4.3.1 K-fold cross validation results

MAE with corresponding standard deviations in the 5-fold cross validation experiment for the MIMIC data are given in Table 3, while the results for the everyday-life data are given in Table 4.

Algorithm	**MAE_{SBP} [mmHg]**
Dummy (predicts mean)	19.70 ± 16.07
Linear regression	18.47 ± 15.91
Ensemble (all feat.)	5.83 ± 7.74
Ensemble (relevant feat.)	4.90 ± 6.59
Algorithm	**MAE_{DBP} [mmHg]**
Dummy (predicts mean)	8.73 ± 6.77
Linear regression	8.14 ± 7.98
Ensemble (all feat.)	2.92 ± 4.09
Ensemble (relevant feat.)	2.21 ± 3.70

Table 3: MAE of different algorithms for SBP and DBP estimation in 5-fold cross validation using the MIMIC hospital dataset.

Algorithm	**MAE_{SBP} [mmHg]**
Dummy (predicts mean)	11.46 ± 7.51
Linear regression	11.21 ± 8.00
Ensemble (all feat.)	9.12 ± 7.90
Ensemble (relevant feat.)	7.87 ± 7.47
Algorithm	**MAE_{DBP} [mmHg]**
Dummy (predicts mean)	5.01 ± 3.99
Linear regression	5.01 ± 8.00
Ensemble (all feat.)	4.38 ± 3.74
Ensemble (relevant feat.)	3.84 ± 3.63

Table 4: Mean absolute errors of different algorithms for SBP and DBP estimation in 5-fold cross validation using the JSI-collected everyday-life dataset.

Ensemble of shallow regression trees has shown the best predictive performance in the 5-fold cross validation for both SBP and DBP using both datasets. We also notice a slightly better performance when only the relevant features, as given by RReliefF, are used in comparison to the default feature set.

As the ensemble of regression trees has shown the best performance, its hyperparameters were optimized using Bayesian optimization. All the available hyperparameters were optimized using the MATLAB built-in Bayesian Optimization Workflow [16]. It optimizes both the hyperparameters of the ensemble as well as the hyperparameters of the weak learners, which are chosen to be shallow Regression Trees. The optimization is ran for 30 iterations, trying to minimize the objective cross-validation loss function. Bootstrap aggregation was chosen as superior over gradient boosting strategy, and the optimal number of weak learners was determined to be 77. The maximum number of splits in the weak learner was determined to be 1, meaning that the regression trees are in fact regression stumps.

4.3.2 Leave-one-subject-out results

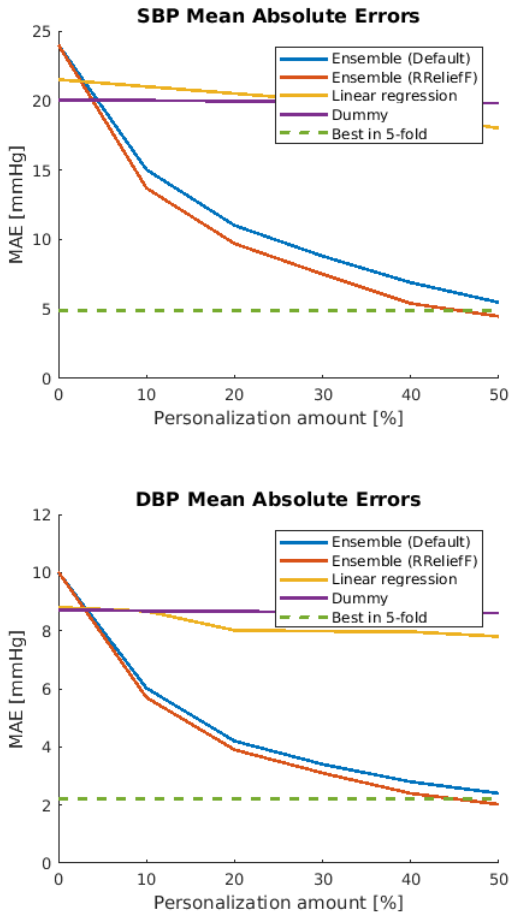

Figure 6: MAE for SBP and DBP for the MIMIC dataset, at different amounts of personalization.

The lowest error using the MIMIC data was again achieved using the hyperparameter tuned Ensemble of regression trees algorithm with RReliefF selected subset of features. The highest amount of personalization (50%) gave the best results. 50% personalization corresponds to 10 BP measurements conducted by the subject, given the fact that 20 segments with 20 different BP values were taken. Obtaining 10 BP measurements by the subject, in order to personalize the model, seems like a reasonable requirement.

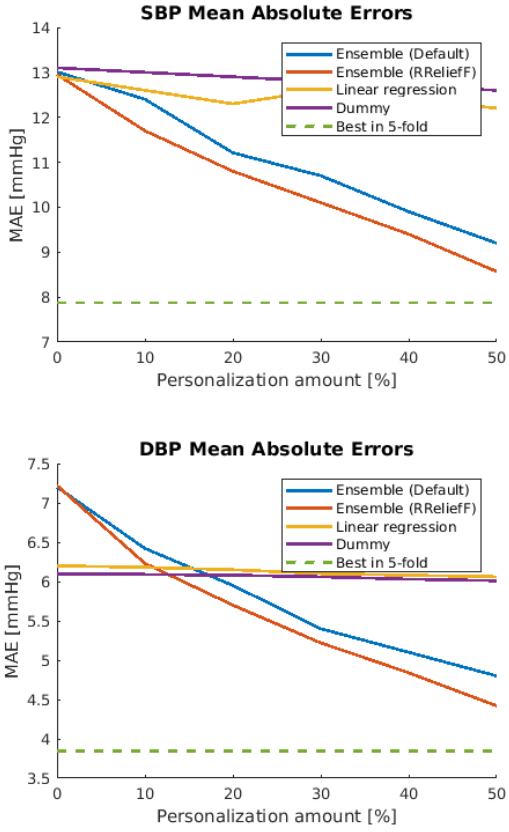

Figure 7: MAE for SBP and DBP for the everyday-life dataset, at different amounts of personalization.

The JSI-collected everyday-life data has proven to be more problematic, as there were only a few different BP values recorded in the first phase of data collection. Furthermore, due to the high amount of movement artefacts, a lot of data was removed by the cleaning algorithm, leaving a very small amount of usable data with a very low variation in BP. This further enhanced the performance of the dummy regressor, which achieved much lower MAE compared to the MIMIC dataset, however, improvements were again achieved by using personalization, as shown in Figure 7.

5 Conclusion

We have developed a system for BP estimation using only the PPG signal, and have evaluated its performance on two distinct datasets using two experimental setups.

The first module of the system deals with signal preprocessing, removing most movement artefacts and anomalies from the PPG signal. It then detects PPG cycles corresponding to heart beats and feeds them to the second module, which computes a number of features describing each cycle. This is followed by feature subset selection using the RReliefF algorithm and finally the features are fed into several regression algorithms. Predictive models were created and evaluated on a hospital MIMIC dataset as well as an everyday-life dataset collected at JSI. The lowest MAE achieved for the MIMIC hospital dataset in 5-fold cross validation were 4.90 $\pm$ 6.59 mmHg for SBP and 2.21 $\pm$ 3.70 mmHg for DBP. The best performing algorithm was an Ensemble of shallow regression trees. Its hyperparameters were optimized using Bayesian optimization. Finally, the same models were evaluated on the same dataset using the leave-one-subject-out validation, achieving the lowest MAE of 4.47 $\pm$ 5.85 mmHg for SBP and 2.02 $\pm$ 2.94 mmHg for DBP, again using the same hyperparameter-tuned Ensemble and the subset of features selected by the RReliefF algorithm. These results were achieved using the maximum, 50% personalization. Similar trends can be observed for the everyday-life JSI-collected dataset. The lowest MAE in 5-fold cross validation were 7.87 $\pm$ 7.47 mmHg for SBP and 3.84 $\pm$ 3.63 mmHg for DBP. Ensemble of shallow regression trees with optimized parameters prevailed again. In LOSO validation, the lowest MAE of of 8.57 $\pm$ 7.93 mmHg for SBP and 4.42 $\pm$ 3.61 mmHg for DBP were achieved.

5.1 Interpretation of results

Comparing the results of the 5-fold cross-validation to those of the LOSO evaluation, we first notice, that the best performing algorithm is the same. In each fold in the 5-fold cross validation, 80% of randomly shuffled instances were taken for training, which translates to 80% personalization for each subject. This is the reason behind the lower MAE in the 5-fold cross-validation, however, similar MAE was also achieved with higher amounts of personalization in the LOSO experiment. The developed system shows promising results and could be used by both regular people and hypertensive patients during their everyday routine, by wearing an unobtrusive wristband. It could inform them of their current medical condition regarding BP. Further testing with more field-collected data is required to more accurately determine its performance, however, it already achieves low MAE when personalization is considered.

5.2 Future work

We plan to expand our data collection experiment at JSI, which will give us more data and more variety within the collected BP data. Once enough data is collected, we plan to upgrade the machine learning part of our pipeline using deep-learning algorithms. These are well-suited for problems dealing with signal analysis and represent the state of the art approach in signal processing in recent years, making them a suitable candidate for our domain.

Acknowledgement

The HeartMan project has received funding from the European Unions Horizon 2020 research and innovation programme under grant agreement No 689660. Project partners are Jožef Stefan Institute, Sapienza University, Ghent University, National Research Council, ATOS Spain SA, SenLab, KU Leuven, MEGA Electronics Ltd and European Heart Network.

References

[1] The World Health Organization. “The top 10 causes of death”, 2015.

[2] Mayo Foundation for Medical Education and Research (MFMER). “Blood pressure chart: What your reading means”. Accessed online: 2nd March, 2018.

[3] Teng et. al. “Continuous and noninvasive estimation of arterial blood pressure using a photoplethysmographic approach”, *Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (IEEE Cat. No.03CH37439)*, 2003.

[4] Frese et al. “Blood Pressure Measurement Guidelines for Physical Therapists”, *Cardiopulmonary Physical Therapy Journal*, 2011.

[5] Shelley et al. “Pulse Oximeter Waveform: Photoelectric Plethysmography”, *Clinical Monitoring: Practical applications for anesthesia and critical care*, 2001.

[6] Tamura et al. “Wearable Photoplethysmographic SensorsPast and Present”, *Electronics*, 2014.

[7] Goldberger et al. “PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals”, *Circulation*, 2000.

[8] Moody et al. “A Database to Support Development and Evaluation of Intelligent Intensive Care Monitoring”, *Computers in Cardiology*, 1996.

[9] Lamonaca et al. “A neural network-based method for continuous blood pressure estimation from a PPG signal”, *IEEE International Congress I2MTC*, 2013.

[10] Lamonaca et al. “Application of the Artificial Neural Network for blood pressure evaluation with smartphones”, *2013 IEEE 7th International Conference on Intelligent Data Acquisition and Advanced Computing Systems (IDAACS)*, 2013.

[11] Najarian et al. “Biomedical Signal and Image Processing, 2nd Edition”, *CRC Press*, 2012.

[12] Robnik-Šikonja et al. “Theoretical and Empirical Analysis of ReliefF and RReliefF”, *Machine Learning*, 2003.

[13] Xing et al. “Optical Blood Pressure Estimation with Photoplethysmography and FFT-Based Neural Networks”, *Biomedical Optics Express*, 2016.

[14] Lzaro et al. “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*, 2014.

[15] Li et al. “Dynamic time warping and machine learning for signal quality assessment of pulsatile signals”, *Physiological Measurement*, 2012.

[16] The MathWorks Inc., Natick, Massachusetts, United States. “MATLAB 2017a Optimization Toolbox”, 2017.

[17] Coleman et al. “Validation of the Omron M7 (HEM-780-E) oscillometric blood pressure monitoring device according to the British Hypertension Society protocol”, *Blood Pressure Monitoring*, 2008.