Assessment of deep learning based blood pressure prediction from PPG and rPPG signals

Fabian Schrumpf
Leipzig University of Applied Sciences
fabian.schrumpf@htwk-leipzig.de

Patrick Frenzel
Leipzig University of Applied Sciences
patrick.frenzel@htwk-leipzig.de

Christoph Aust
University Hospital Leipzig
christoph.aust@googlemail.com

Georg Osterhoff
University Hospital Leipzig
georg.osterhoff@medizin.uni-leipzig.de

Mirco Fuchs
Leipzig University of Applied Sciences
mirco.fuchs@htwk-leipzig.de

Abstract

Exploiting photoplethysmography signals (PPG) for non-invasive blood pressure (BP) measurement is interesting for various reasons. First, PPG can easily be measured using fingerclip sensors. Second, camera-based approaches allow to derive remote PPG (rPPG) signals similar to PPG and therefore provide the opportunity for non-invasive measurements of BP. Various methods relying on machine learning techniques have recently been published. Performances are often reported as the mean average error (MAE) on the data which is problematic. This work aims to analyze the PPG- and rPPG-based BP prediction error with respect to the underlying data distribution. First, we train established neural network (NN) architectures and derive an appropriate parameterization of input segments drawn from continuous PPG signals. Second, we apply this parameterization to a larger PPG dataset and train NNs to predict BP. The resulting prediction errors increase towards less frequent BP values. Third, we use transfer learning to train the NNs for rPPG based BP prediction. The resulting performances are similar to the PPG-only case. Finally, we apply a personalization technique and retrain our NNs with subject-specific data. This slightly reduces the prediction errors.

1. Introduction

Blood pressure (BP) is regarded as an essential biomarker for various diseases. Techniques for discontinuous measurements are quite elaborated and comprise auscultatory and oscillometric cuff-based methods. While they are commonly used in both clinical and home environments, they are rather unsuited for long-term measurements due to patient discomfort and potential skin irritations. Techniques for continuous BP monitoring are also readily available for some use cases and comprise arterial BP measurement and cuff-less sensor solutions [1, 2, 3]. The former is invasive thus limited to clinical settings, the latter requires the use of multiple sensors, e.g. ECG electrodes and PPG sensors. They actually allow continuous monitoring but are still uncomfortable for patients. Long-term measurements even require regular recalibration using an additional cuff. Other but usually less practicable techniques to measure BP comprise ultrasound, tactile sensor-based approaches, and vascular unloading-based methods [4].

In recent years research mainly focused on BP estimation from cuff-less multi and single sensor solutions. The former often utilize time or phase differences between different signals (usually ECG and PPG or multiple PPG) related to the blood volume propagation through arteries [5, 6, 7, 8, 9]. The latter mainly exploit morphological properties of blood volume dynamics derived from PPG measurements on a particular site [10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20]. PPG-only based methods are particularly interesting. Not only do they target single sensor solutions, but rather is their underlying signal generation principle very much similar to that of camera-based techniques known as rPPG. If PPG signals could be utilized for BP estimation, a fully contactless method like rPPG might be feasible as well, thus providing opportunities for a lot of clinical and non-clinical application scenarios. In fact, some studies have investigated the estimation of BP [21, 22, 23] or its correlates [24, 25, 26, 27] from these signals.

Recent progress is driven by machine learning (ML) techniques and recently by the advent of deep learning methods. They typically pursue feature-based approaches or perform learning in an end-to-end manner. Feature-based methods exploit spectral or temporal properties of the PPG signal which are fed into a learning algorithm to predict systolic and diastolic BP. End-to-end procedures leverage the waveforms themselves and implicitly derive features to predict BP. Often the accuracies of these methods were reported to be in line with well-established standards (BHS or AAMI) [28]. It is, however, common to only report an indicator for the mean performance of an algorithm without sufficiently taking the underlying data distributions into account. This is problematic since the performance of a learning algorithm tends to be biased towards the mode of this distribution. That means the error appears to be small for the majority of samples but is much larger when deviating towards the tails of their distribution. To the best of our knowledge, only one recent work studied the distribution of the BP prediction error [29]. They reported an insufficient accuracy of the investigated method with respect to the full BP range but only divided the BP range coarsely into three intervals. This underpins that a more detailed analysis of the error distribution is critical for the assessment of BP prediction methods in particular for clinical applications.

The BP distribution of a data set is affected by various issues. The variability of the pulse morphology among subjects, e.g. caused by age and cardiovascular diseases, certainly affects the association of the signal shape to a particular BP. The equipment in a hospital setting differs as well and, more importantly, the contact pressure of a PPG sensor can affect the pulse morphology [30, 31, 32]. Therefore the ability of the learning algorithm to generalize well may be impaired. On the individual scale of a subject, the BP variation is often limited during the measurement, e.g. due to medication (often not reported for databases), limited physical activity or due to short record periods. This may also affect the data distribution, particularly when training, validation and test sets are not split carefully. Lastly, the data used to train a learning model is often not publicly available. The question which subjects were used for training and testing when using publicly available databases remains frequently unanswered. Hence, it is challenging to assess whether an improved prediction error results from methodological improvements or just from data selection. While these are issues that arise for both PPG and rPPG based methods, recent works disputed the usefulness of rPPG signals for BP estimation altogether [33, 34].

This paper targets (1) an empirical evaluation of the parameterization of input signals (e.g. segment cropping from the continuous signals) suited for both PPG and rPPG based BP prediction with established neural network (NN) architectures; (2) a detailed assessment of PPG based BP prediction performance on sufficiently small intervals of the systolic and diastolic targets; (3) rPPG based BP prediction on a data set recorded in a clinical setting based on a pretrained and fine-tuned network; (4) the effect of personalization by fine-tuning networks using subject-specific data.

2. Related work

2.1. Deriving BP using parameterized models

Early works for cuffless BP estimation utilize the pulse transit (PTT) or pulse arrival time (PAT). The PTT is the time delay for the pulse to travel between two different arterial sites and the PAT is the time delay between the electrical onset (R-peak in the ECG) and the arrival of the evoked pulse wave at a particular site [35]. They can be used to derive the pulse wave velocity (PWV) using the Moens-Korteweg equation [36, 37]. Gesche *et al*. used the PWV and inferred BP values by means of a linear regression model [38]. This method resulted in a commercially available smartwatch for sleep research (SOMNOtouch NIBP, SOMNOMEDICS GmbH). Socrates *et al*. [9] validated this device and found a standard error of 4.2 mmHg for systolic and diastolic BP during sleep and awake phases.

2.2. BP prediction using PPG features

Haddad *et al*. used morphological features of the PPG-waveform to classify BP into normal and hypertonic ranges using multi-linear regression [18]. Others employed dense neural networks (DNN) to derive BP from time-based morphological features such as pulse width, pulse amplitudes and heart rate. They also extracted features from the first and second derivative of the PPG waveform [13, 15, 19]. Other authors used recurrent neural networks (RNN) to derive BP from time- and frequency-based PPG-features [5, 14, 39]. In [6, 7], the authors trained a very deep RNN by introducing skip connections between layers to overcome the vanishing gradient problem [40]. Yang *et al*. constructed time series from morphological PPG features and PTT values. They divided the time series into high- and low-frequency components and fed them into a DNN and RNN for BP estimation [8].

2.3. End-to-End approaches to predict BP

Slapničar *et al*. [10] used a parallel architecture consisting of three residual neural networks (ResNet) to predict BP using the PPG waveform and its first and second derivatives. They also used a subject-based calibration resulting in a significantly reduced mean average error (MAE). Schlesinger *et al*. used a siamese convolutional neural network (CNN) to predict BP variations with respect to a calibration value. They used PPG spectrograms as inputs for their NN [12]. Baek *et al*. utilized ECG and PPG to derive time and fre-

quency domain input segments and proposed a new BP prediction model based on multiple losses[41].

Recent studies [42, 43, 44] used CNNs in combination with RNNs to estimate BP. The features vectors are derived from PPG based CNN embeddings and fed into an RNN for BP prediction. Eom *et al*. [43] achieved an MAE of 0.06 mmHg and 5.42 mmHg for systolic and diastolic BP using this approach. Wang *et al*. used the PPG time-course and first and second-order derivatives as input to a parallel structure consisting of an LSTM and a CNN. They classified BP into five different classes and achieved an overall accuracy of 91% [16]. A more straightforward approach was pursued by Han *et al*. who used a CNN with PPG-waveforms as input and achieved an overall F1-score of 0.9 when classifying BP into different hypertension classes [20]. Xing *et al*. used a simple DNN for BP estimation. They transformed the PPG signal into the frequency domain and used its amplitude and phase spectrum as input. The mean error based on the MIMIC-II database was 0.06 mmHg and 0.01 mmHg for systolic and diastolic BP [11].

2.4. BP estimation from rPPG

Camera-based estimation of BP has gained significant attention in recent years. Most works derive the pulse wave from image sequences of a subject's face from single or multiple regions. Some methods derive rPPG just by averaging of the facial skin pixels in the green or red channel [21, 24, 25]. More sophisticated approaches account for light reflections and movements to reduce artefacts [22, 23]. These rPPG signals are subsequently used to estimate BP using supervised machine learning methods.

Luo *et al*. used transdermal imaging to derive the pulse wave from 17 different regions of interest (ROI) in the face [22]. Amplitude and time domain features were fed into a multilayer perceptron. They achieved a mean error of 0.39 mmHg and -0.2 mmHg for systolic and diastolic BP. Jain *et al*. derived the pulse wave using principal component analysis (PCA) on the red color channel. Time and frequency features of the pulse signal are used to predict BP with linear regression model. Their model achieved an MAE of 3.9 mmHg and 3.7 mmHg for systolic and diastolic BP [21]. Other approaches translate the pulse wave velocity and image-based pulse transit time (iPTT) concept to remote measurements. In [24] and [27], iPTT values measured between the palm and the face region revealed a high correlation to PTT values based on ECG and PPG. The authors also found a moderate correlation between the iPPT values and the BP values taken by a cuff device [24]. Many authors studied the iPTT between different ROIs in the face but only found low correlations to BP [25, 26]. In [25], largest correlations were found for the subject's right palm, corroborating the findings of [24, 27]. Tran *et al*. used the iPTT between the face and the palm to predict BP using a

pre-trained NN. Their MAE for systolic and diastolic BP was 3.1 mmHg and 2.6 mmHg [23].

3. Methods

3.1. Datasets

3.1.1 PPG data

The MIMIC-III database consists of thousands of records from various hospitals collected between 2001 and 2008 and sampled at 125 Hz [45]. We used a subset of the MIMIC-III database available on Kaggle[1]. It consists of 12000 records of PPG, ECG and ABP signals. Their authors applied extensive preprocessing on the waveforms to provide a clean and valid dataset [46, 47]. It represents only a small fraction of the MIMIC-III database. but is therefore compact and contains signals with acceptable quality. However, the subject affiliation is unknown thus rendering it unsuited for evaluating the model performance. Hence, we only used this dataset in the first part of our work to evaluate the parameterization of input signals. It is denoted as MIMIC-A for the remainder of this paper.

A much larger portion of the MIMIC-III database was downloaded using scripts provided by [10], resulting in a total of 4000 records (PPG and ABP signal pairs). This dataset is denoted as MIMIC-B and is used for performance evaluations.

3.1.2 rPPG data

Data for camera-based BP prediction was recorded in a study at the Leipzig University Hospital. The study design was approved by the ethics committee of the University of Leipzig. Subjects were informed about the content of the study and gave written consent prior to taking part. 50 subjects with a planned surgical intervention were enrolled. After surgery, the patients were transferred to the intensive care unit. Our recording system consisted of an industrial USB camera (IDS UI-3040CP, 32 fps) connected to a PC (Intel NUC NUC7I7BNH). Videos of the subjects face and upper body with approximately two hours duration were recorded. Ground truth BP was derived from the bedside monitor with one minute temporal resolution.

3.2. Neural network architectures

We used three different neural network architectures to predict BP values. The first is AlexNet which is a CNN architecture originally developed for image classification [48]. We adopted its structure in order to use PPG time series as inputs and return systolic and diastolic BP values instead of class predictions.

---

[1]https://www.kaggle.com/mkachuee/BloodPressureDataset

Second, we used a ResNet, i.e. a very deep CNN architecture [40]. Their skip connections efficiently account for the vanishing gradient problem occuring in deep architectures. The ResNet was modified in the same manner as the AlexNet. The dimensions for the input layer of these networks were $N_{samp} \times 1$ in the univariate case and $N_{samp} \times 3$ when using the raw time series and its first and second order derivatives. The final classification in each original model was replaced by a regression layer consisting of two neurons for SBP and DBP with a linear activation function.

Third, we used the architecture published by Slapničar *et al.* [10]. It consists of a spectrotemporal network with residual connections. It is a parallel structure and processes PPG signals and their first and second-order derivatives.

3.3. PPG signal processing

We first aimed at studying the effect of the signal length on the prediction performance. We divided the PPG and ABP signals from the MIMIC-A dataset into segments of different lengths, i.e. 1, 2, 5, 7, 9, 11, 13, 15, 17 and 20s. This dataset is called *const_time_xx*. Since this approach leads to an interruption of single PPG cycles at the beginning and the end of each time window, we created an alternative dataset to evaluate whether this affects the prediction performance. Therefore, we ensured that only full cycles (i.e. complete beats) are contained in a time window under test using the following procedure. We estimated the heart rate based on the PPG by detecting the spectral component with the highest amplitude. Next, we divided the signal into time segments containing an integer number of 1, 2, 5, 7, 9, 11, 13, 15, 17 and 20 PPG-waves. All segments were resampled to have equal length. It is obvious that this also eliminates absolute temporal information. The new sampling frequency was chosen in a way that the PPG-window contains a heart rate of 60 bpm. The resulting dataset is called *const_beats_xx*. We processed ABP signals in a similar manner to yield datasets consisting of PPG-ABP pairs.

Recent studies showed that, apart from the raw PPG itself, derivatives yield useful information on the cardiovascular state [49] and can be useful for BP as well [10, 15, 16]. Hence, we employed the first and second order derivatives of each PPG-window as well and studied whether this multivariate approach reduces the BP prediction error.

We derived the ground truth systolic and diastolic BP from the ABP segments. We used a peak detection algorithm to detect systolic and diastolic peaks [50, 51] and derived the reference BP as the median of all peaks within each segment. We employed several plausibility checks. All BP values outside a physiologically plausible range of 75 to 165 mmHg and 40 to 80 mmHg for systolic and diastolic BP, respectively, were discarded. Median heart rates in each window that exceeded the ranges of 50 to 140 bpm were also rejected.

3.4. Evaluation of NN input sequences

We first trained our NNs with the MIMIC-A dataset to determine the proper cropping strategy and window length. We used 100000 randomly drawn samples for training the model and additional 25000 samples for validating and testing, respectively. The neural architectures and the training pipelines were implemented using Google Tensorflow 2.4 and Python 3.5 (Adam optimizer, $\alpha = 0.001$, 50 epochs, mean squared error loss). The models with the lowest validation loss were used for subsequent performance tests.

We conducted three repeated training procedures for both AlexNet and ResNet for each length of the input time segment under test. We then employed a paired t-test to evaluate whether interrupting PPG cycles when cropping segments from the continuous PPG signal affects the MAE. We conducted this analysis separately for the AlexNet and ResNet architecture.

To determine the optimal window length we employed a twofold strategy. First, we evaluated the NN performances with respect to the input length of the segments derived from PPG signals. Second, we evaluated how a particular length would affect the SNR of rPPG segments. We aimed at selecting a length that, on the one hand, ensures high SNR values (thus resulting in longer segments) but, on the other hand, maximizes the number of segments (and therefore enforcing shorter segments) available for NN training. The latter is motivated by the fact that various studies have shown that a large number of training examples is crucial for the successful training of NNs. Hence, our procedure aimed at finding a trade-off resulting in a sufficiently large number of training examples with an acceptable SNR for rPPG, given the limited amount of samples in the rPPG dataset. As before, we divided the rPPG dataset into windows of different length and calculated the SNR for every time window. We used a threshold of 7 dB SNR above which a rPPG segment would be accepted. We finally used a the shortest window length that provides a large number of acceptable time windows while still being long enough to expect good results for BP prediction with the NN.

3.5. PPG-based BP prediction

3.5.1 Data preprocessing

The MIMIC-B dataset was divided into windows using the optimal cropping strategy and window length determined in section 3.3. Since this data came directly from the physionet.org database, additional preprocessing steps had to be applied. First, a 4th order Butterworth band-pass filter was applied to the PPG-signals. Cut-off frequencies were set to 0.5 Hz and 8 Hz. Second, the signal-to-noise ratio (SNR) was calculated for all signals [52]. All signal windows with an SNR below −7 dB were discarded. All PPG-windows were normalized to zero mean and unit variance.
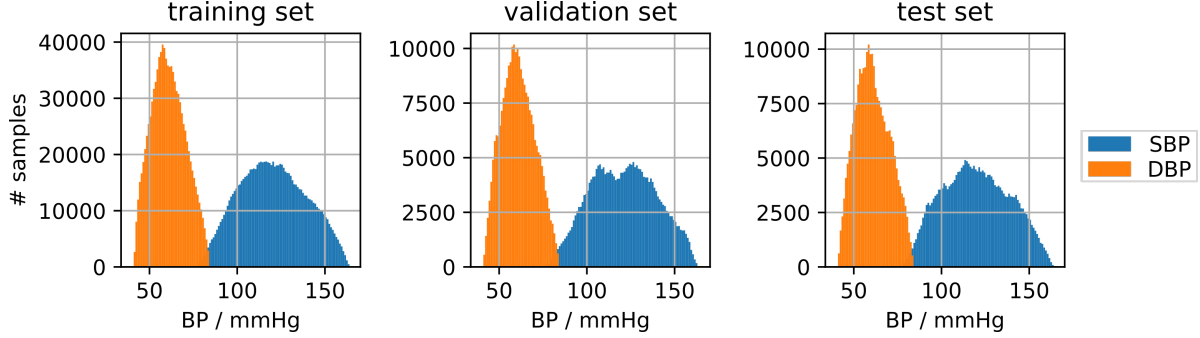
Figure 1. Blood pressure distribution in the training, validation and test set in dataset MIMIC-III-B based on the MIMIC-III database

3.5.2 NN training and evaluation

The dataset was split into training, validation and test set on a subject-basis to prevent contamination of the validation and test set by training data. We used 3750 subjects for training and 625 subjects for validation and testing. Among these subjects, we have randomly drawn 1.000.000 samples for training, 250.000 samples for validation and 250.000 samples for testing. The BP distribution among these datasets can be seen in Fig. 1.

Input pipelines and NNs (AlexNet, ResNet, model from Slapničar *et al.*) were implemented using Google Tensor-Flow 2.4 and Python 3.8 was used for training (Adam optimizer, $\alpha = 0.001$, euclidian loss, 60 epochs). We used the models with the lowest MAE on the validation set for further testing.

For evaluation purposes, we additionally used a mean regressor that always predicts the systolic and diastolic BP from the training set. A well generalizing ML method will exceed the mean regressor's performance. We used the MAE metric to asses the performance of all methods. In contrast to other work, we determined the prediction errors both for the full dataset and separately for rather narrow BP bins, altogether spanning the range of ground truth BP values contained in the datasets.

3.6. rPPG-based BP prediction

3.6.1 Preprocessing

ROIs on a subject's forehead and cheeks were labelled manually after the recording using custom software. We used the Plane-orthogonal-to-skin (POS) algorithm to derive the pulse wave from the skin pixels [53]. The resulting rPPG-signal was then inspected visually. Twenty-five subjects with heavy motion artifacts, frequent movement or insufficient lighting were deemed unsuitable and excluded from further analysis. Remaining data was divided into windows based on their heart rate. Seven heartbeats were included (compare Sec. 3.4 and 4.1.1) and each window was resampled according to the procedure for the MIMIC-B dataset (Sec. 3.5.1). We calculated the SNR value for each rPPG

window according to [52] and excluded windows with an SNR below $-7\,\mathrm{dB}$. Ground truth BP values were downloaded from the bedside monitor.

3.6.2 Transfer learning

The resulting dataset was used to fine-tune the pre-trained NNs. Transfer learning exploits the idea that rPPG and PPG waveforms share similar properties and should therefore give rise to similar relevant features during NN training. Due to the low amount of data, however, an entire retraining of these NNs was not feasible. We therefore only optimized the final layer while freezing all other network weights. We used the Adam optimizer ($\alpha = 0.001$) and fine-tuned until the MAE stopped decreasing. The best model given by leave-one-out cross-validation was then used to evaluate the model using the test subject. We also investigated whether the personalization strategy presented in [10] improves the prediction accuracy. In addition to the training dataset we reserved 20% of the test subject's data for training and validated the model with the remaining 80%.
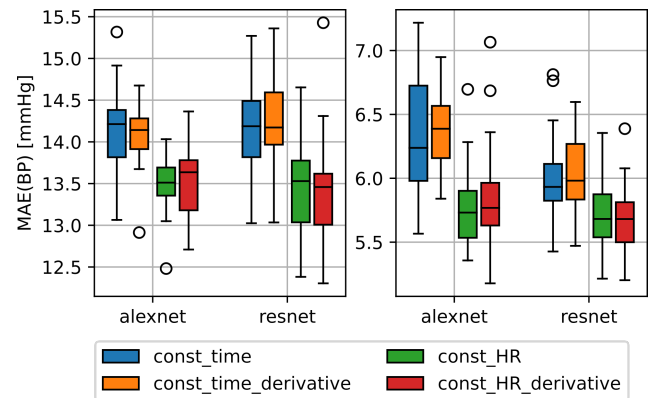

Figure 2. Difference in MAE between *const_time* and *const_HR* datasets for AlexNet and ResNet. (a): MAE for the systolic blood pressure; (b): MAE for diastolic blood pressure; *const_time_derivative* and *const_HR_derivative* datasets include the first and second order derivatives of the PPG windows contained in the datasets *const_time* and *const_HR*.
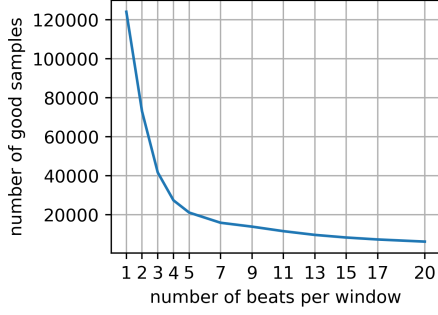
Figure 3. Fraction of acceptable samples in the rPPG dataset with respect to the window length. Samples with an SNR below the threshold of -7 dB were discarded.

4. Results

4.1. PPG-based BP prediction

4.1.1 Input signals

Figure 2 shows the training results of the AlexNet and ResNet architectures with the *const_HR* and *const_time* datasets with and without additional time derivatives. Note that we combined the results obtained for all variations of the segment lenghts for this analysis. The prediction errors based on *const_HR* are lower in comparison to *const_time*. The significance of these findings was confirmed using a paired t-Test ($p < 0.01$). The use of the first and second order derivatives shown as *const_HR_derivative* and *const_time_derivative* did not yield to a general improvement compared to the univariate case. Just the SBP MAE of the AlexNet was slightly lower than in the univariate case. Due to this and for the sake of simplicity, we did not consider derivatives for our further analysis.

According to Sec. 3.4, we aimed to derive a suitable length of the input signal as a trade-off from PPG and rPPG data. First, we evaluated the prediction error with respect PPG input signals. We expected a slight increase of the error towards longer segments since the morphology of PPG cycles slightly varies over time and therefore would introduce undesired ambiguities with respect to the underlying BP. However, our empirical analysis (3 repetitions for each NN and length parameter) did not confirm this effect and resulted in an almost equal prediction error for each tested length. Note that because of the high computational effort that would be necessary to obtain a sufficiently high number of repeated training procedures for each NN and length parameter (30+ each), a statistical justification of this effect can not be provided. From this perspective, the longest possible segment length (20 s) seemed useful.

Second, we analyzed the SNR of rPPG segments with respect to their lengths (Fig. 3) and aimed to maximize the number of resulting samples available for training. It should be noted that the SNR measure becomes less appropriate to-
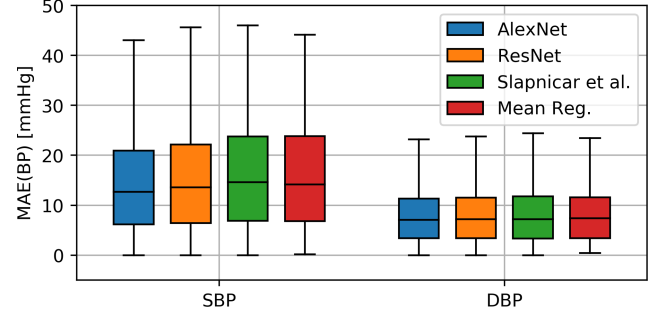

Figure 4. Comparison of the MAE for SBP and DBP of the three analyzed CNN architectures in comparison to the mean regressor.

wards small segment lengths which also explains the sharp decline at the beginning of the curve. It seemed reasonable to select a value beyond the bend of the curve to account for this effect. We finally selected 7 s as the segment length for all further analyses.

4.1.2 Predicting BP using PPG data

Results after training the CNN architectures with the bigger MIMIC-B dataset can be seen in Fig. 4. It is evident that the DBP MAE is lower than the SBP MAE. This can be attributed to the fact that its range of values is smaller than the SBP's range. We compared the MAE of the AlexNet, ResNet and the model from Slapničar *et al.* to the MAE of the mean regressor using a paired t-Test. Figure 4 shows that particularly the DBP errors of the NNs were very similar to those of the mean regressor. However, the DBP MAE
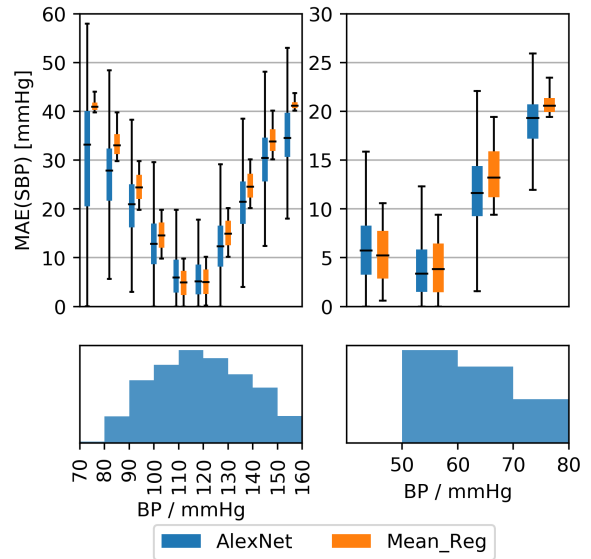

Figure 5. MAE depending on the ABP-based blood pressure for the AlexNet architecture. The admissible blood pressure range was divided into bins of width 10 mmHg. For reference, the distribution of SBP and DBP is displayed at the bottom of the plot.
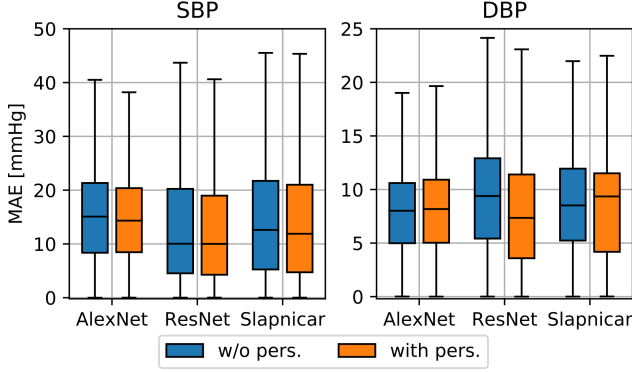
Figure 6. Comparison of the MAE between fine tuning the neural networks with and without personalization.

	SBP [mmHg]		DBP [mmHg]	
	Mean	Std	Mean	Std
no personalization				
AlexNet	15.7	9.66	8.27	4.67
ResNet	13.02	10.55	9.81	9.88
Slapničar *et al*.	14.04	9.82	8.64	4.77
Mean Reg.	13.4	10.1	8.9	5.2
with personalization				
AlexNet	15.2	9.11	8.52	4.92
ResNet	12.51	12.61	8.3	9.84
Slapničar *et al*.	13.56	10.12	8.5	5.02
Mean Reg.	13.9	10.4	8.5	5.0

Table 1. MAE of the neural networks after fine-tuning the pre-trained networks using rPPG-data. Top: Fine-tuning *without* personalization; Bottom: Fine-tuning *with* personalization using 20% of the test subject's data for training.

of the AlexNet and ResNet architectures were significantly lower than the mean regressor. This could not be found for the model Slapničar *et al*. When analyzing the SBP MAE, we discovered that AlexNet and ResNet achieved significantly lower errors than the mean regressor.

To analyze the BP in dependence of the MAE we divided the acceptable BP input range into bins of $10\,\mathrm{mmHg}$ width and calculated the error for each bin separately for each architecture. Since the results are very similar, only those for the AlexNet are shown (Fig. 5). The error strongly varies across BP bins thus emphasizing a large dependence of the error on the underlying BP. Each architecture achieves the lowest MAE in a range of 100 - 120 mmHg (SBP) and 56 - 60 mmHg (DBP). The BP distribution of the test set is shown at the bottom of Fig. 5. It can be seen that the error is inversely proportional to the number of samples in the respective bin. This suggests that the ML methods achieve a training loss reduction by concentrating their predictions on the BP range with the most frequent training examples.

4.2. rPPG based BP prediction

Next, we fine-tuned the NNs trained on PPG data for BP prediction using rPPG data. From 14 subjects included in the analysis, 12 were used for training, 1 for validation and 1 for testing.

Table 1 (top) shows the overall MAE and STD after fine-tuning with rPPG data without personalization. It can be seen that the ResNet model achieves the lowest MAE in terms of SBP. AlexNet achieves the lowest DBP MAE. The results of the model from Slapničar *et al*. are close to the mean regressor. Figure 6 (blue boxes) shows a boxplot of the MAE based on rPPG signals. The differences in SBP MAE and DBP MAE are statistically significant ($p \leq 0.05$). This emphasizes that the results may be very different on a per-subject basis but the overall performance of the three NNs is quite similar both among each other and in comparison to the mean regressor.

Finally, we compared the impact of personalization on

the BP prediction error after fine-tuning the NN. Figure 6 (red boxes) and Tab. 1 (bottom) show that personalization has the biggest effect on the AlexNet and ResNet architectures. The SBP MAE improved by 0.5 mmHg and 0.51 mmHg, respectively. The DBP MAE of the ResNet improved by 1.29 mmHg. The SBP MAE of the Slapničar *et al*. network improved by 0.48 mmHg. However, changing the training/test split also influences the performance of the mean regressor. Its SBP MAE increased by 0.5 mmHg while the DBP MAE decreased by 0.4 mmHg. Hence, the model from Slapničar *et al*. performed slightly better than the mean regressor without personalization and is on par with the mean regressor after personalization.

5. Discussion

This paper investigated the feasibility of BP estimation using PPG- and rPPG-based pulse wave signals. Importantly, our aim was not to derive a particularly accurate model to achieve state-of-the-art performance for BP prediction from PPG but rather (1) to explore the dependence of these models on some important time domain properties of the PPG and rPPG input signal, (2) to learn how these models perform not only in terms of a mean performance on a given dataset but on a more fine grained scale of multiple BP bins and (3) to investigate the feasibility of transfer learning from PPG for rPPG based BP prediction. We also ensured a careful split of the data and divided the dataset based on subjects to avoid contaminating the validation and test data with training data.

First, we conducted an empirical evaluation of the parameterization of the input signals that were suited to train our NNs. In particular, we analyzed the window length, cropping of segments from continuous signals and the use of derivatives. We used established NN architectures (i.e., AlexNet and ResNet) from the literature and adopted them

for BP prediction. Moreover, we used the architecture as presented by Slapničar *et al.* which is optimized for BP prediction from PPG. We found that the use of derivatives does not provide significant improvements and is less important. Avoiding the introduction of strong phase discontinuities using an appropriate cropping of segments from continuous signals significantly improves the prediction performance. Finally, the total length of the input sequence was less important with respect to the PPG-based prediction errors. The selection of the length of segments was mainly driven by rPPG data to maximize the number samples in the dataset.

Second, our analysis of the BP dependent prediction error reveals that the NNs are partly superior over the mean regressor. However, effect sizes were small. Besides that, we found a strong dependence of the bin-wise prediction error on the number samples in the particular BP bin in the underlying distribution. The most accurate predictions occurred in BP bins containing the most samples. This emphasizes the tendency towards predicting the mode of the training dataset. A similar dependence was found in [29], which led to the retraction of a publicly available smartphone app for BP prediction. However, their results are based on a much coarser subdivision of the BP range.

In order to compare our results to previous work, we additionally evaluated the mean performance. None of the results met the requirements as defined in the relevant BHS and AAMI standards, which require the probability of a BP measurement device to provide an acceptable error ($BP \leq 10\,\mathrm{mmHg}$) to exceed 85% [28]. Especially the high MAEs in the higher and lower BP ranges pose a problem since many clinical applications rely on an acceptable accuracy in hypo- and hypertensive ranges. Importantly, our BP MAE is in accordance with Slapničar *et al.* [10] who also accounted for subject specific affiliations to training and testing datasets. Note that we did not employ any hyperparameter tuning which might further improve our results, but certainly only on a gradual scale. In contrast, other authors who did not explicitly mention a subject-based dataset split reported substantially lower prediction errors [30, 31, 32, 54]. These differences suggest that morphological inter-individual variations due to age, comorbidities, medication and measurement equipment prevent the investigated NNs to generalize well. Recent works by Zhang *et al.* addressed this issue using neural architectures specifically tailored to learn domain-invariant features [55]. This was beyond the scope of our study since we focused on NN architectures which are already applied on a broad basis.

Our study emphasizes the following: To develop an ML technique that aims to be of relevance for practical clinical applications, one must (1) evaluate the prediction error of a model over the full systolic and diastolic BP range and (2) therefore carefully take the data distributions in training and test sets into account. While this is of course obvious and an important rule for designing ML algorithms in general, it is not yet treated with the necessary care among published literature for ML based BP prediction. In particular, differences in age, the health state and medications are likely to have a strong influence on the BP distribution in a dataset. It is also worth to further investigate the relation between the PPG morphology and the sensor contact pressure [30] which might have an even more severe effect on the ability of an ML model to generalize at all. This is especially the case when using public databases where patient records can emerge from various sources. Given our findings, it does not seem unlikely that the mentioned and probably even more issues causing variations in the PPG morphology might render BP predictions from PPG a highly ill-posed problem for real world applications. An appropriate split into subject-dependent training and test sets must be ensured in any case. Including data of training subjects in the test set by randomly splitting the datasets on a sample-basis leads to an overestimation of the model performance.

Third, we fine-tuned our NNs using rPPG data conducted in a clinical study. We found that the estimation error greatly varied between subjects. Given the ResNet, the overall error was still smaller than the mean regressor. Like for PPG, we also investigated the effect of personalization. We discovered a considerably improved prediction accuracy in comparison to the mean regressor. We emphasize that we had only limited training data for fine-tuning and therefore could only fine-tune the final layer of each NN. Given more data, it would be conceivable to tune additional layers and possibly enhance the prediction performance.

Remote measurement of BP using standard RGB cameras is still an active field of research. While it still seems questionable to derive BP from PPG-only data, it remains even more questionable whether rPPG data is actually suitable for BP estimation [33, 34, 56]. Future studies should concentrate on end-to-end approaches since the rPPG-signal's low SNR hampers methods based on morphological features, especially when different skin tones, motion and changes in illumination are involved [57, 58].

Acknowledgements

References

[1] X.-R. Ding, N. Zhao, G.-Z. Yang, R. I. Pettigrew, B. Lo, F. Miao, Y. Li, J. Liu, and Y.-T. Zhang, "Continuous Blood Pressure Measurement From Invasive to Unobtrusive: Celebration of 200th Birth Anniversary of Carl Ludwig," *IEEE Journal of Biomedical and Health Informatics*, vol. 20, no. 6, pp. 1455–1465, Nov. 2016. 1

[2] R. Kumar, P. Dubey, A. Zafer, A. Kumar, and S. Yadav, "Past, Present and Future of Blood Pressure Measuring Instruments and Their Calibration," *Measurement*, p. 108845, Dec. 2020. 1

[3] T. Tamura, "Cuffless blood pressure monitors: Principles, standards and approval for medical use," *IEICE Transactions on Communications*, 2020. 1

[4] R. Mukherjee, S. Ghosh, B. Gupta, and T. Chakravarty, "A Literature Review on Current and Proposed Technologies of Noninvasive Blood Pressure Measurement," *Telemedicine and e-Health*, vol. 24, no. 3, pp. 185–193, Mar. 2018. 1

[5] Y.-H. Li, L. N. Harfiya, K. Purwandari, and Y.-D. Lin, "Real-Time Cuffless Continuous Blood Pressure Estimation Using Deep Learning Model," *Sensors*, vol. 20, no. 19, p. 5606, Sep. 2020. 1, 2

[6] M. S. Tanveer and M. K. Hasan, "Cuffless blood pressure estimation from electrocardiogram and photoplethysmogram using waveform based ANN-LSTM network," *Biomedical Signal Processing and Control*, vol. 51, pp. 382–392, May 2019. 1, 2

[7] P. Su, X.-R. Ding, Y.-T. Zhang, J. Liu, F. Miao, and N. Zhao, "Long-term blood pressure prediction with deep recurrent neural networks," in *2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*. Las Vegas, NV, USA: IEEE, Mar. 2018, pp. 323–328. 1, 2

[8] S. Yang, J. Sohn, S. Lee, J. Lee, and H. C. Kim, "Estimation and Validation of Arterial Blood Pressure using Photoplethysmogram Morphology Features in conjunction with Pulse Arrival Time in Large Open Databases," *IEEE Journal of Biomedical and Health Informatics*, pp. 1–1, 2020. 1, 2

[9] T. Socrates, P. Krisai, A. S. Vischer, A. Meienberg, M. Mayr, and T. Burkard, "Improved agreement and diagnostic accuracy of a cuffless 24-h blood pressure measurement device in clinical practice," *Scientific Reports*, vol. 11, no. 1, p. 1143, Dec. 2021. 1, 2

[10] G. Slapničar, N. Mlakar, and M. Luštrek, "Blood Pressure Estimation from Photoplethysmogram Using a Spectro-Temporal Deep Neural Network," *Sensors*, vol. 19, no. 15, p. 3420, Aug. 2019. 1, 2, 3, 4, 5, 8

[11] X. Xing and M. Sun, "Optical blood pressure estimation with photoplethysmography and FFT-based neural networks." *Biomedical optics express*, vol. 7, no. 8, pp. 3007–20, Aug. 2016. 1, 3

[12] O. Schlesinger, N. Vigderhouse, D. Eytan, and Y. Moshe, "Blood Pressure Estimation From PPG Signals Using Convolutional Neural Networks And Siamese Network," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Barcelona, Spain: IEEE, May 2020, pp. 1135–1139. 1, 2

[13] Y. Kurylyak, F. Lamonaca, and 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)*. Minneapolis, MN, USA: IEEE, May 2013, pp. 280–283. 1, 2

[14] C. El Hajj and P. A. Kyriacou, "Cuffless and Continuous Blood Pressure Estimation From PPG Signals Using Recurrent Neural Networks," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. Montreal, QC, Canada: IEEE, Jul. 2020, pp. 4269–4272. 1, 2

[15] Y.-C. Hsu, Y.-H. Li, C.-C. Chang, and L. N. Harfiya, "Generalized Deep Neural Network Model for Cuffless Blood Pressure Estimation with Photoplethysmogram Signal Only," *Sensors*, vol. 20, no. 19, p. 5668, Oct. 2020. 1, 2, 4

[16] D. Wang, X. Yang, X. Liu, S. Fang, L. Ma, and L. Li, "Photoplethysmography based stratification of blood pressure using multi information fusion artificial neural network," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Seattle, WA, USA: IEEE, Jun. 2020, pp. 1113–1119. 1, 3, 4

[17] A. Dash, N. Ghosh, A. Patra, and A. D. Choudhury, "Estimation of Arterial Blood Pressure Waveform from Photoplethysmogram Signal using Linear Transfer Function Approach," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. Montreal, QC, Canada: IEEE, 2020, pp. 2691–2694. 1

[18] S. Haddad, A. Boukhayma, and A. Caizzone, "Continuous PPG-Based Blood Pressure Monitoring Using Multi-Linear Regression," *arXiv:2011.02231 [physics]*, Nov. 2020, arXiv: 2011.02231. 1, 2

[19] R. K. Pandey, T.-Y. Lin, and P. C.-P. Chao, "Design and implementation of a photoplethysmography acquisition system with an optimized artificial neural network for accurate blood pressure measurement," *Microsystem Technologies*, Jan. 2021. 1, 2

[20] C. Han, M. Gu, F. Yu, R. Huang, X. Huang, and L. Cui, "Calibration-free Blood Pressure Assessment Using An Integrated Deep Learning Method," p. 5. 1, 3

[21] M. Jain, S. Deb, and A. V. Subramanyam, "Face video based touchless blood pressure and heart rate estimation," in *2016 IEEE 18th International Workshop on Multimedia Signal Processing (MMSP)*. Montreal, QC, Canada: IEEE, Sep. 2016, pp. 1–5. 1, 3

[22] H. Luo, D. Yang, A. Barszczyk, N. Vempala, J. Wei, S. J. Wu, P. P. Zheng, G. Fu, K. Lee, and Z.-P. Feng, "Smartphone-Based Blood Pressure Measurement Using Transdermal Optical Imaging Technology," *Circulation: Cardiovascular Imaging*, vol. 12, no. 8, pp. 1–10, Aug. 2019. 1, 3

[23] Q.-V. Tran, S.-F. Su, Q.-M. Tran, and V. Truong, "Intelligent Non-Invasive Vital Signs Estimation From Image Analysis," in *2020 International Conference on System Science and Engineering (ICSSE)*. Kagawa, Japan: IEEE, Aug. 2020, pp. 1–6. 1, 3

[24] I. C. Jeong and J. Finkelstein, "Introducing Contactless Blood Pressure Assessment Using a High Speed Video Camera," *Journal of Medical Systems*, vol. 40, no. 4, p. 77, Apr. 2016. 1, 3

[25] N. Sugita, M. Yoshizawa, M. Abe, A. Tanaka, N. Homma, and T. Yambe, "Contactless Technique for Measuring Blood-Pressure Variability from One Region in Video Plethysmography," *Journal of Medical and Biological Engineering*, vol. 39, no. 1, pp. 76–85, Feb. 2019. 1, 3

[26] R. Takahashi, K. Ogawa-Ochiai, and N. Tsumura, "Non-contact method of blood pressure estimation using only facial video," *Artificial Life and Robotics*, Jul. 2020. 1, 3

[27] J. Finkelstein and I. c. Jeong, "Towards Contactless Monitoring of Blood Pressure at Rest and During Exercise Using Infrared Imaging," in *2020 11th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*. New York City, NY: IEEE, Oct. 2020, pp. 0756–0758. 1, 3

[28] G. S. Stergiou, B. Alpert, S. Mieke, R. Asmar, N. Atkins, S. Eckert, G. Frick, B. Friedman, T. Graßl, T. Ichikawa, J. P. Ioannidis, P. Lacy, R. McManus, A. Murray, M. Myers, P. Palatini, G. Parati, D. Quinn, J. Sarkis, A. Shennan, T. Usuda, J. Wang, C. O. Wu, and E. O'Brien, "A Universal Standard for the Validation of Blood Pressure Measuring Devices: Association for the Advancement of Medical Instrumentation/European Society of Hypertension/International Organization for Standardization (AAMI/ESH/ISO) Collaboration Statement," *Hypertension*, vol. 71, no. 3, pp. 368–374, Mar. 2018. 2, 8

[29] M. Dörr, S. Weber, R. Birkemeyer, L. Leonardi, C. Winterhalder, C. J. Raichle, N. Brasier, T. Burkard, and J. Eckstein, "iPhone App compared with standard blood pressure measurement –The iPARR trial," *American Heart Journal*, p. S0002870320304026, Dec. 2020. 2, 8

[30] A. Chandrasekhar, M. Yavarimanesh, K. Natarajan, J.-O. Hahn, and R. Mukkamala, "PPG Sensor Contact Pressure Should Be Taken Into Account for Cuff-Less Blood Pressure Measurement," *IEEE Transactions on Biomedical Engineering*, vol. 67, no. 11, pp. 3134–3140, Nov. 2020. 2, 8

[31] J. Allen, J. O'Sullivan, G. Stansby, and A. Murray, "Age-related changes in pulse risetime measured by multi-site photoplethysmography," *Physiological Measurement*, vol. 41, no. 7, p. 074001, Aug. 2020. 2, 8

[32] J. Přibil, A. Přibilová, and I. Frollo, "Comparative Measurement of the PPG Signal on Different Human Body Positions by Sensors Working in Reflection and Transmission Modes," *Engineering Proceedings*, vol. 2, no. 1, p. 69, Nov. 2020. 2, 8

[33] A. A. Kamshilin, E. Nippolainen, I. S. Sidorov, P. V. Vasilev, N. P. Erofeev, N. P. Podolian, and R. V. Romashko, "A new look at the essence of the imaging photoplethysmography." *Scientific reports*, vol. 5, p. 10494, May 2015, publisher: Nature Publishing Group. 2, 8

[34] A. Moco, S. Stuijk, M. van Gastel, and G. de Haan, "Impairing Factors in Remote-PPG Pulse Transit Time Measurements on the Face," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, vol. 2018-June. IEEE, Jun. 2018, pp. 1439–14 398, issue: June ISSN: 21607516. 2, 8

[35] G. Zhang, M. Gao, D. Xu, N. B. Olivier, and R. Mukkamala, "Pulse arrival time is not an adequate surrogate for pulse transit time as a marker of blood pressure," *Journal of Applied Physiology*, vol. 111, no. 6, pp. 1681–1686, Dec. 2011. 2

[36] C. F. Wippermann, D. Schranz, and R. G. Huth, "Evaluation of the pulse wave arrival time as a marker for blood pressure changes in critically ill infants and children," *Journal of Clinical Monitoring*, vol. 11, no. 5, pp. 324–328, Sep. 1995. 2

[37] F. J. Callaghan, C. F. Babbs, J. D. Bourland, and L. A. Geddes, "The relationship between arterial pulse-wave velocity and pulse frequency at different pressures," *Journal of Medical Engineering & Technology*, vol. 8, no. 1, pp. 15–18, Jan. 1984. 2

[38] H. Gesche, D. Grosskurth, G. Küchler, and A. Patzak, "Continuous blood pressure measurement by using the pulse transit time: Comparison to a cuff-based method," *European Journal of Applied Physiology*, vol. 112, no. 1, pp. 309–315, Jan. 2012, publisher: Springer ISBN: 1439-6327 (Electronic)\r1439-6319 (Linking). 2

[39] U. Senturk, K. Polat, and I. Yucedag, "A non-invasive continuous cuffless blood pressure estimation using dynamic Recurrent Neural Networks," *Applied Acoustics*, vol. 170, p. 107534, Dec. 2020. 2

[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA: IEEE, Jun. 2016, pp. 770–778. 2, 4

[41] S. Baek, J. Jang, and S. Yoon, "End-to-End Blood Pressure Prediction via Fully Convolutional Networks," *IEEE Access*, vol. 7, pp. 185 458–185 468, 2019. 3

[42] J. Esmaelpoor, M. H. Moradi, and A. Kadkhodamohammadi, "A multistage deep neural network model for blood pressure estimation using photoplethysmogram signals," *Computers in Biology and Medicine*, vol. 120, p. 103719, May 2020. 3

[43] H. Eom, D. Lee, S. Han, Y. S. Hariyani, Y. Lim, I. Sohn, K. Park, and C. Park, "End-To-End Deep Learning Architecture for Continuous Blood Pressure Estimation Using Attention Mechanism," *Sensors*, vol. 20, no. 8, p. 2338, Apr. 2020. 3

[44] D. U. Jeong and K. M. Lim, "Combined Deep CNN–LSTM Network-based Multitasking Learning Architecture for Noninvasive Continuous Blood Pressure Estimation using Difference in ECG-PPG Features," In Review, preprint, Jan. 2021. 3

[45] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark, "MIMIC-III, a freely accessible critical care database," *Scientific Data*, vol. 3, no. 1, p. 160035, Dec. 2016. 3

[46] M. Kachuee, M. M. Kiani, H. Mohammadzade, and M. Shabany, "Cuff-less high-accuracy calibration-free blood pressure estimation using pulse transit time," in *2015 IEEE International Symposium on Circuits and Systems (ISCAS)*. Lisbon, Portugal: IEEE, May 2015, pp. 1006–1009. 3

[47] ——, "Cuffless Blood Pressure Estimation Algorithms for Continuous Health-Care Monitoring," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 4, pp. 859–869, Apr. 2017. 3

[48] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, May 2017. 3

[49] M. Elgendi, R. Fletcher, Y. Liang, N. Howard, N. H. Lovell, D. Abbott, K. Lim, and R. Ward, "The use of photoplethysmography for assessing hypertension," *npj Digital Medicine*, vol. 2, no. 1, p. 60, Dec. 2019. 4

[50] M. Elgendi, "Detection of c, d, and e waves in the acceleration photoplethysmogram," *Computer Methods and Programs in Biomedicine*, vol. 117, no. 2, pp. 125–136, Nov. 2014. 4

[51] M. Elgendi, I. Norton, M. Brearley, D. Abbott, and D. Schuurmans, "Systolic Peak Detection in Acceleration Photoplethysmograms Measured from Emergency Responders in Tropical Conditions," *PLoS ONE*, vol. 8, no. 10, p. e76585, Oct. 2013. 4

[52] G. de Haan and A. van Leest, "Improved motion robustness of remote-PPG by using the blood volume pulse signature," *Physiological Measurement*, vol. 35, no. 9, pp. 1913–1926, Sep. 2014. 4, 5

[53] W. Wang, A. C. den Brinker, S. Stuijk, and G. de Haan, "Algorithmic Principles of Remote PPG," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 7, pp. 1479–1491, Jul. 2017. 5

[54] C. M. McEniery, Yasmin, I. R. Hall, A. Qasem, I. B. Wilkinson, and J. R. Cockcroft, "Normal Vascular Aging: Differential Effects on Wave Reflection and Aortic Pulse Wave Velocity," *Journal of the American College of Cardiology*, vol. 46, no. 9, pp. 1753–1760, Nov. 2005. 8

[55] L. Zhang, N. C. Hurley, B. Ibrahim, E. Spatz, H. M. Krumholz, R. Jafari, and B. J. Mortazavi, "Developing Personalized Models of Blood Pressure Estimation from Wearable Sensors Data Using Minimally-trained Domain Adversarial Neural Networks," *arXiv:2007.12802 [cs]*, Jul. 2020, arXiv: 2007.12802. 8

[56] A. V. Moço, S. Stuijk, and G. de Haan, "New insights into the origin of remote PPG signals in visible light and infrared," *Scientific Reports*, vol. 8, no. 1, p. 8501, Dec. 2018. 8

[57] ——, "Skin inhomogeneity as a source of error in remote PPG-imaging," *Biomedical Optics Express*, vol. 7, no. 11, p. 4718, Nov. 2016. 8

[58] Y.-C. Lin and Y.-H. Lin, "A study of color illumination effect on the SNR of rPPG signals," in *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. Seogwipo: IEEE, Jul. 2017, pp. 4301–4304. 8