Regression or Classification? Reflection on BP prediction from PPG data using Deep Neural Networks in the scope of practical applications

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

Paul Rudi Serdack
Leipzig University of Applied Sciences
paul_rudi.serdack@stud.htwk-leipzig.de

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

Abstract

Photoplethysmographic (PPG) signals offer diagnostic potential beyond heart rate analysis or blood oxygen level monitoring. In the recent past, research focused extensively on non-invasive PPG-based approaches to blood pressure (BP) estimation. These approaches can be subdivided into regression and classification methods. The latter assign PPG signals to predefined BP intervals that represent clinically relevant ranges. The former predict systolic (SBP) and diastolic (DBP) BP as continuous variables and are of particular interest to the research community. However, the reported accuracies of BP regression methods vary widely among publications with some authors even questioning the feasibility of PPG-based BP regression altogether. In our work, we compare BP regression and classification approaches. We argue that BP classification might provide diagnostic value that is equivalent to regression in many clinically relevant scenarios while being similar or even superior in terms of performance. We compare several established neural architectures using publicly available PPG data for SBP regression and classification with and without personalization using subject-specific data. We found that classification and regression models perform similar before personalization. However, after personalization, the accuracy of classification based methods outperformed regression approaches. We conclude that BP classification might be preferable over BP regression in certain scenarios where a coarser segmentation of the BP range is sufficient.

*This research was funded in part by the German Federal Ministry of Economics and Technology (BMWi) (FKZ 49VF170043). The study at the Leipzig University Medical Center was conducted according to the guidelines of the Declaration of Helsinki, and approved by the Ethics Committee of the University of Leipzig (protocol code: 170/19-ek; date of approval: 6 July 2019).

1. Introduction

Predicting BP from single sensor signals such as PPG has gained a lot of attraction in recent years [5, 23]. Using PPG is particularly interesting not only because sensors are cheap and easy to apply but also because the technique is related to remote PPG (rPPG; or imaging PPG / iPPG). RPPG generally refers to camera/video based derivation of PPG signals which allows to conduct remote measurements without any physical contact. If a reliable prediction of BP from PPG would be possible, there is reason to belief that concepts could be transferred and expanded to rPPG based BP prediction, which in fact has already been approached in several studies [30, 34, 40].

From a machine learning perspective, current approaches that use PPG for BP prediction can broadly be categorized into approaches based on extracting hand-crafted features [6, 10, 35] and approaches that employ the entire signal and sometimes also its derivatives [25, 29]. The latter are usually based on a certain deep neural network (DNN) architecture. In DNN, these signals or their spectral representations are then usually either used directly in an end-to-end learning scheme to predict BP from the shape information [25] or are transformed into a spectrogram beforehand [38]. Even hybrid approaches have already been investigated [3, 29].

With respect to the target variable, approaches can be categorized into BP classification and BP regression. Classification is usually restricted to scenarios where authors are interested in predicting hypotension or hypertension versus normal BP [4, 33]. Regression, on the other hand, predicts BP as a continuous variable. With respect to the evaluation, usually only a mean error for the whole dataset along with a standard deviation or a some related metric is reported [14, 20, 22]. Some of these results imply suitability of the reported methods for medical applications since the

errors reported are well below or at least close to the requirements of medical standards [3, 14]. In fact, a recent study has shown that much larger errors must be expected for a wide range of DNN architectures if the true underlying BP is not within the normotensive BP range [27].

Assessing the reported results is often not straightforward. A reliable conclusion is usually difficult mainly due to issues with the underlying dataset. Main problems are (1) the rather unbalanced nature of the underlying datasets which might lead to overfitting towards the mean of the distribution [27], (2) issues with personalization within the datasets rendering the methods prone to overfitting to individual subjects [17] and (3) the use of proprietary data collected in an experimental setting where the BP usually lacks the necessary variation within its range of interest. These varying boundary conditions render comparisons among different studies difficult especially if those studies try to solve different problems (i.e. BP classification and BP regression).

It stands to reason that well performing regression methods would be particularly interesting for medical applications. This would however require a robust performance within the full range of physiologically meaningful BP values. On the other hand, some applications certainly do not need highly accurately predicted BP values but rather a reliable mapping to a certain BP interval which we will refer to as BP bin for the rest of this paper (e.g. [4,33]). The method of dividing the BP range into suitable intervals (i.e. number and widths of the intervals) to achieve adequate accuracy for clinical purposes is still an open question.

It is conceivable that certain scenarios only require a binary classification (high vs. low BP). From a medical perspective a meaningful segmentation would be the subdivision into intervals representing hypotension (hypoT), normotension (NT) and hypertension (HT) resulting in a multi class classification approach. On the other end, a dense binning could consist of very narrow BP intervals (e.g. bins of width 1 mmHg). It is obvious that the latter case is more suited for a regression-based approach since the underlying sorted nature of the BP values exhibits a larger importance with respect to the predicted target variable compared to cases with only a small number of classes.

Given these very different perspectives on the underlying problem, the following questions can be raised : (i) Is there reason to believe that classification and regression approaches for BP prediction should be preferred one over another in certain scenarios? (ii) Is there a trade-off between approaching the problem using a classification approach or a regression approach and the desired discretization of the BP range of interest into bins? (iii) And more generally, are classification approaches perhaps generally more suitable for BP prediction with respect to realistic scenarios since they might provide a better generalization to the available datasets which come with the above mentioned issues?

In this study, we compare regression and classification approaches based on several DNN architectures for BP prediction. From an evaluation perspective and for comparison reasons, we treat the whole problem as a classification task. While we directly derive a predicted class in a classification approach, we manually assign the target variable to a predicted class in the regression approaches based on its location in a certain BP interval. We applied this scheme to several segmentations of the BP range (i.e. subdivision with differently sized BP bins/intervals). Given each segmentation, we carefully prepared datasets based on the MIMIC-III database which allowed us to prevent overfitting to individual subjects as well as to any of the incorporated BP intervals. That means we ensured that our training sets are balanced with respect to the particular range segmentation into multiple intervals.

Other studies reported a positive effect of personalizing the network prior to prediction with subject-specific data [13,20,29]. This means that the pretrained network is partly or completely retrained using some portion of data from the subject used for prediction. Following this idea we also investigated in this study whether this personalization scheme is more or less effective with respect to the segmentation of the target variable (BP range) in the underlying BP prediction task. There is reason to believe that personalization is particularly effective for narrow bins, i.e. in problems very much related to regression but becomes less important if only a low number of BP classes with wide intervals are used.

The original contributions of this paper are as follows. First, we provide a systematic comparison of classification and regression based approaches for the task of BP classification using several BP range segmentations. Second, we evaluate the effect of personalization of DNNs using subject-specific data with respect to the granularity of the segmentation of the BP range for both the regression and classification scenario. Moreover, we show that the findings hold true for several DNN architectures.

The remainder of this paper is as follows: Section 2 gives an overview over existing work in the field of BP prediction and classification. Section 3 outlines the methods used for creating the dataset, describes the relevant neural architectures and how they were trained in each scenario as well as the performance measures used for evaluating the results. Section 4 presents the results and draws comparisons between regression-based and classification based approach with and without personalization. Section 5 assesses the results presented in the light of our initial hypotheses. We also draw conclusions as to if and under what circumstances a decision for or against regression or classification might be justified.

2. Related Works

2.1. BP regression

Existing literature regarding PPG-based BP prediction can be divided into approaches that try to predict SBP as well as DBP as continuous variables (i.e. regression) and approaches that aim at classifying PPG signals into a set of predefined classes (i.e. classification). The definition of these classes is usually targeted towards the detection of certain adverse medical conditions (e.g. hypotension (HypoT), normortension (NT), hypertension (HT)). Since training of very deep neural architectures became feasible on consumer-grade hardware, the attention of the research community has shifted mainly towards regression-based approaches.

Regression-based approaches can be further subdivided into methods using parameterized models or PPG features as well as end-to-end approaches. Parameterized methods use the pulse transit time (PTT) or pulse arrival time (PAT) to derive the pulse wave velocity (PWV). Several studies used linear regression models to infer BP values from PWV [7, 37, 39]

Feature-based approaches derive time and frequency domain features from PPG signals and use them as input to neural networks (NNs). Recently, Mahmud et al. [22] designed a U-Net architecture that was tasked with deriving informative features from PPG and electrocardiography (ECG) signals. These features were then used to predict SBP and DBP using several clustering algorithms (e.g. kNN approaches and support vector machines). They achieved a MAE of 2.33 mmHg for SBP and 0.713 mmHg for DBP, respectively. Other authors designed recurrent NNs and derived time- and frequency-based features from PPG and ECG for predicting BP [6, 11, 26, 28].

In contrast to feature-based methods, end-to-end approaches leverage the PPG waveforms themselves and implicitly derive informative features to predict BP. The advantage of this approach is that the selection and derivation of specific features is not necessary leaving the task of detecting patterns in the input data that are correlated to BP entirely to the neural architecture. Aguet et al. [2] trained a siamese NN. They averaged consecutive PPG time windows to improve feature robustness to create the datasets for training and testing. Together with a callibration measurement the siamese network achieved an mean error of 0.31 mmHg for SBP and 0.4 mmHg for DBP, respectively. However, the standard deviation of these errors was high (10.27 mmHg for SBP and 5.62 mmHg for DBP). Leitner et al. [20] designed a hybrid neural architecture consisting of convolutional, recurrent and fully connected layers. After training with PPG signals, the NN was fine tuned using additional data from the test subjects. The authors achieved an MAE of 3.52 mmHg (SBP) and 2.2 mmHg (DBP). Jeong et al. [14] used a convolutional neural network (CNN) in combination with a recurrent NN to predict BP. They processed PPG end ECG from the MIMIC-III database to create the dataset used for training and testing. Their architecture achieved a mean error of 0.02 mmHg (SBP) and 0.16 mmHg (DBP).

2.2. BP classification

Many recent publications aiming at BP classification use already established neural architectures that are successfully applied to image classification tasks. Cano et al. [4] modified pretrained GoogleNet and ResNet architectures and trained them using 50 subjects downloaded from the MIMIC-III database. The target variable was a classification into HT, pre-HT and NT. The highest F1-score on the test dataset was achieved with the ResNet18 network. Sun et al. [33] used the Hilbert-Huang Transform on PPG signals and their first and second derivatives to fine tune a pretrained AlexNet architecture. The classification of the input signals into NT and HT achieved an accuracy score of 98.9 %. Liang et al. [21] computed the continuous wavelet transform (CWT) from the PPG signals of 121 records downloaded from the MIMIC-III database. A pretrained GoogLeNet was used to classify the scalograms into NT, pre-HT and HT. Multiple trainings were performed to invesitgate the accuracy when classifying the scalograms into pairwise combinations of the target classes. The highest F1-score of 92.55 % was achieved when classifying NT and pre-HT.

Other authors created custom neural architectures for BP classification. Wu et al. [38] proposed a CNN designed for NT /HT classification based on the CWT of PPG signals. They achieved a validation accuracy of 90 %. Mejía-Mejía et al. [24] derived time and frequency domain features from the PPG-based pulse variablity signal. A subset among those features was selected based on a importance analysis. The authors trained various classification methods (e.g. k-NN, support vector machines and multilayer perceptrons) for HypoT/NT/HT classification. The highest accuracy on the test set was 70 %.

3. Methods

3.1. Dataset

The dataset used in this work is based on the MIMIC-III database [9, 15, 16] and was created as described in [27]. Essential processing steps comprised downloading ABP and PPG signals from 4000 subjects off the MIMIC-III database using mining scripts provided by Slapničar et al. [29]. The downloaded ABP and PPG data were divided into windows with a length of 5 s and an overlap of 2.5 s (50 %) between consecutive windows. To create a balanced dataset, the samples had to be collected so that each subject contributed

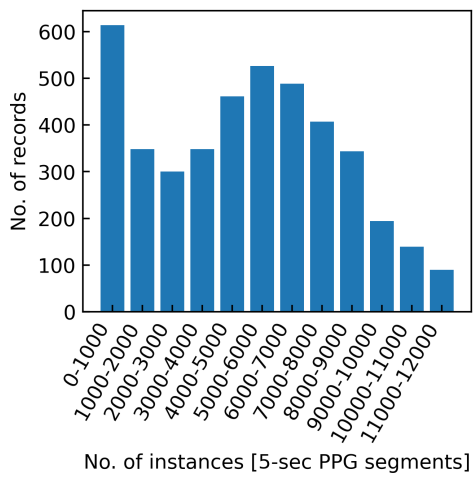

Figure 1. Histogram of the number of records downloaded from the MIMIC-III database containing a certain amount of PPG windows for training

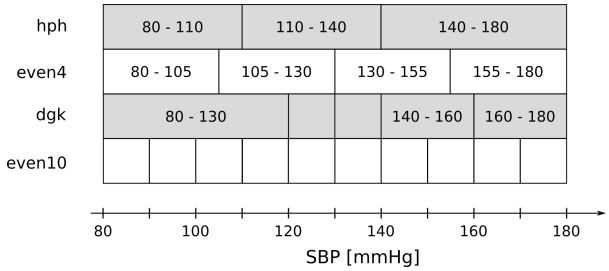

Figure 2. Overview over the BP range segmentation used for training the NNs. The SBP range was divided into variable numbers of bins.

BP range segmentation	No of subj,	No of samples
hph	1214	3642000
even4	475	1900000
dgk	189	1134000
even10	94	940000

Table 1. Numbers of subjects and records included in the segmentations of the BP range used to train NNs. For details on the segmentations, see text.

equally to the dataset. Figure 1 shows a histogram of the number of records in the MIMIC-III database that contain a certain amount of samples. Since the majority of the records contribute at least 1000 samples to the dataset, only those records were selected for further processing. PPG signals were filtered using a 4th order Butterworth bandpass filter ($f_{cut} = 0.5Hz - 8Hz$). Additionally, a quality check in terms of signal-to-noise ratio (SNR) was performed for every PPG window. Samples with an SNR below -7 dB were discarded. All PPG windows were normalized to zero mean and unit variance.

Ground truth SBP values were derived from the ABP signals. We selected the SBP as the sole target since it has proven to be a better indicator for cardiovascular risk than DBP [32]. We employed a peak detection algorithm to detect the systolic peaks in each ABP window. The reference SBP was then derived by calculating the median of all SBP peaks in each signal window. To discard samples with a BP value outside the physiologically plausible range, we removed all samples with an SBP lower than 80 mmHg or higher than 180 mmHg. Signal windows with a median heart rate that exceeded the range of 50 to 140 bpm were also rejected.

3.2. BP range segmentations

To investigate the influence of different segmentations of the BP range on the classification accuracy, we divided the SBP range into bins. Four different BP segmentation schemes shown in Figure 2 were used. (1) *hph*: 3 bins, their location and width reflects diagnostically meaningful BP ranges (i.e. hypoT, NT and HT) according to the German Cardiac Society (DGK) and the World Health Organization (WHO) [1, 36]; (2) *even4*: four equally sized bins covering the whole BP range; (3) *dgk*: a division of the BP range into six physiologically meaningful intervals according to the DGK [36]; (4) *even10*: 10 bins of width 10 mmHg, this approximates more regression-like approaches. For each BP segmentation we selected only subjects from the MIMIC-III database, whose BP value ranges spanned across all BP bins. In order to create balanced datasets, the contribution of each subject was limited to 1000 samples per bin. This led to datasets containing a variable number of subjects and samples depending on the number of bins in the dataset. An overview of the total number of subjects and samples for every BP segmentation can be found in Table 1.

3.3. Neural network architectures

Four different NN architectures were used for classification and regression. We used a modified version of the AlexNet architecture to classify BP into various bins [18]. Originally, AlexNet is a CNN that takes RGB images as input and classifies them into one of 1000 categories. We adopted the architecture for BP classification such that the first layer takes PPG time series data as input. Similarly, the number of output neurons of the final classification layer was adjusted according to the respective BP segmentation that was used for BP classification.

ResNets are very deep CNN consisting of blocks of convolutional layers with skip connections [12]. These skip connections efficiently account for the vanishing gradient problem that occurs in very deep neural architectures [31].

We used three different versions of this architecture with varying depth. Specifically, we used ResNet18, ResNet34 and ResNet50.

The input dimensions of all networks were $N_{samp} \times 1$ (1D PPG signal segment). Output dimensions of the final classification layers were $N_{bin} \times 1$ for the classification problem, where N_{bin} is given by the number of bins in the particular BP range segmentation. In case of regression based-training the target variable corresponds to the SBP as a single value. Network weights of all models were initialized randomly using the Glorot method since it has proven to lead to a quicker convergence of the NN during training [8].

3.4. NN training

3.4.1 Pretraining

The data was split into chunks of 70 %, 22.5 % and 7.5 % which were used for NN training, validation after each epoch and final testing. The split was achieved by assigning data from each subject to only one of these chunks. In comparison to a data split based on raw data samples this strategy prevents overfitting to subjects as data of particular subjects are not split across training, validation and test sets. Our approach to prevent an imbalance between the number of samples in the BP bins of the particular BP range segmentation is described in sec. 2.2.

Input and training pipelines as well as the neural architectures were implemented using Tensorflow 2.3 and Python 3.8. Training was performed using the Adam optimizer with an initial learning rate of 0.001 and a batch size of 128. The training was stopped if the validation loss stopped improving for 10 epochs and the best performing model was used for testing.

3.4.2 Personalization

Previous studies have shown that there are substantial differences in PPG morphology between different subjects that prevent a successful generalization of NNs across various patients [29]. Consequently, the prediction accuracy of NNs trained on data from an extensive subject population might be inadequate for clinical applications. Using subject-specific data to fine tune the pretrained NN has the potential to increase the prediction accuracy for individual subjects. To investigate the effect of personalization on the BP prediction accuracy for classification and regression based approaches we selected 10 patients from the MIMIC-III dataset that were previously used for testing the NNs. Each chosen subject's data was ordered by SBP value and every 10th sample was selected as data for fine tuning the NN. Since we ensured that the data of every subject spanned the whole SBP range during dataset creation, the data used for finetuning also fulfilled this criterion.

	AlexNet	ResNet18	ResNet34	ResNet50
Class.				
hph	0.45	0.44	0.45	0.44
even4	0.36	0.36	0.37	0.36
dkg	0.24	0.24	0.25	0.23
even10	0.16	0.15	0.16	0.16
Reg.				
hph	0.42	0.46	0.45	0.45
even4	0.36	0.36	0.37	0.38
dkg	0.25	0.25	0.25	0.25
even10	0.14	0.16	0.16	0.16

Table 2. Test accuracy of the prediction performance of the NNs for the different BP range segmentations under test. Results are presented separately for regression-based and classification-based approaches. Predicted SBP values from the regression-based approach were assigned to their respective BP bin to allow for a direct comparison to the classification-based approach in terms of accuracy.

Our fine tuning strategy corresponds to a transfer learning approach [20] in which all weights of a pretrained network are allowed to be updated. The remaining data of the particular subject was split into equal parts and used for validating and testing the fine tuned NNs. Training was performed for a fixed number of 100 epochs. The model from the best epoch in terms of validation accuracy was used for testing.

3.5. Evaluation metric

We evaluated the performance of each NN in terms of accuracy with which the models predicted the correct BP bin. The accuracy metric is well suited since we ensured a balanced number of samples across classes. In the case of the BP regression, the predicted SBP value was assigned to its respective BP bin. This allowed a direct comparison between classification and regression evaluation results using classification-based metrics. Additionally, confusion matrices were calculated for a better grasp of the prediction characteristics.

4. Results

Table 2 shows the accuracy for every BP segmentation and every neural architecture on the test set after training the models from scratch. It can be seen that BP segmentations with just a small number of classes (*hph*) achieve a higher accuracy than BP segmentations that divide the BP range into a bigger number of classes (*dgk*). Comparing the different neural architectures, no significant difference in accuracy could be observed. Likewise, there was no difference in accuracy between the regression-based and the classification-based training.

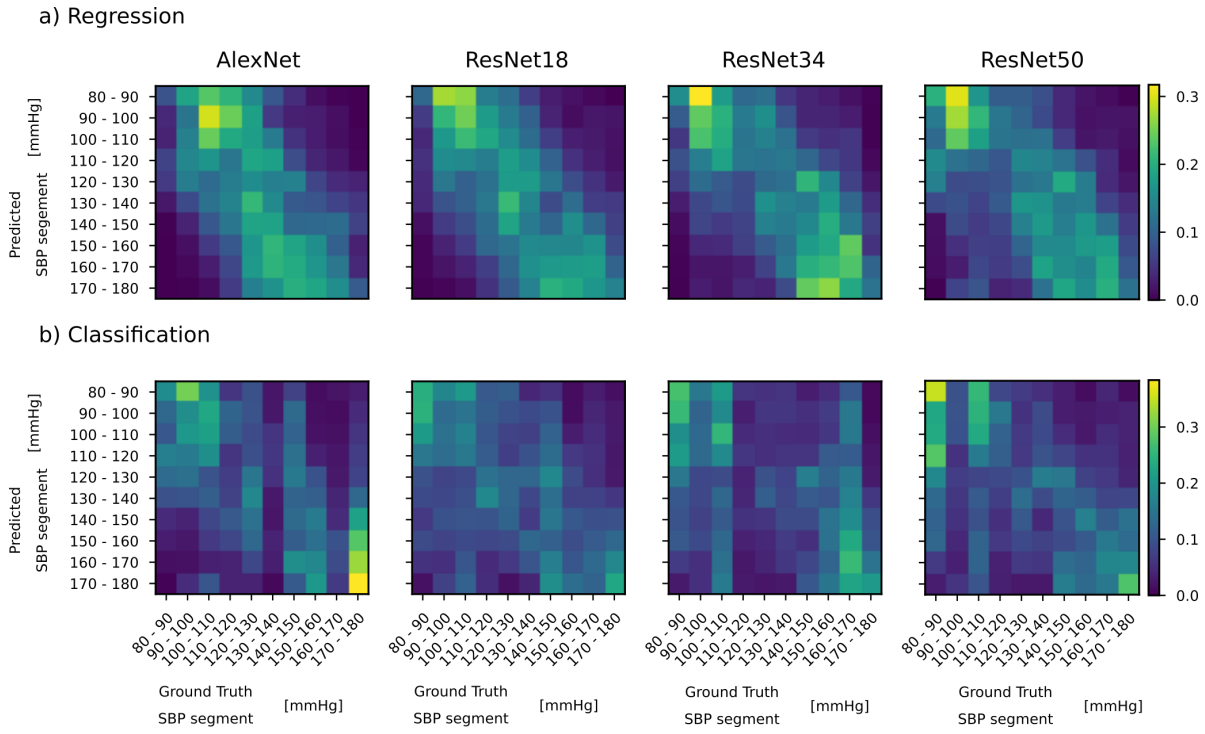

Figure 3. Confusion matrices for the pretraining of the neural architectures using the *even10* segmentation. Ground truth BP classes and predicted BP classes are shown along the x- and y-axis respectively. Results are presented separately for regression-based (a) and classification-based (b) approaches. Number of samples in matrix element are normalized to the total number samples in the respective matrix row (i.e. number of samples in the respective class).

Figure 3 shows the confusion matrices for the regression-based and classification-based approaches using the *even10* segmentation. The false detections indicated in the off-diagonal elements underpin the findings from the accuracy metrics in Table 2. However, resulting matrices from the regression-based approaches show that most of the samples are grouped near the main diagonal suggesting that the prediction is in many cases only off by a few bins from the correct bin. This effect is most pronounced in the results from the *even10* BP segmentation. The models seem to be more capable of distinguishing BP bins that are farther apart in the BP range than bins that are close to each other in comparison to classification models. This is not reflected in the accuracy measure which penalizes misclassification regardless of the proximity of the misclassified bin to the correct bin.

4.1. Personalization

Personalization was performed independently for each of the 10 selected subjects using 10 % of the data. The remaining data was used in equal parts for validation and testing. We additionally verified the test performance on the original test set without the subject used for fine-tuning. This was done to ensure that the generalization capabilities of the network are maintained. Although the results are not shown here, we did not observe a significant drop of the performance after personalization. Figure 4 shows the mean test accuracy for every training scenario. The models were evaluated before (blue bars) and after fine-tuning (orange bars). Test accuracy increased after fine tuning for every training scenario proving the efficacy of personalization for improving the performance on individual subjects. Similar to the results presented in Table 2 as well as Figure 3 the test accuracy declined depending on the number of intervals used for the segmentation of the BP range. BP range segmentations with more bins resulted in a lower accuracy on the test set both pre- and post-personalization in comparison to segmentations with fewer bins. When comparing the various neural architectures, it can be seen that personalization had a greater effect on the ResNet variants as the increase in accuracy post-personalization seems to be higher in comparison to the AlexNet architecture.

Furthermore, we analyzed the differences in test accuracy between classification-based and regression based approaches for every BP range segmentation. Results are depicted in Figure 5. We observed only small differences in pre-personalization accuracy when comparing classification- and regression-based accuracy. However,

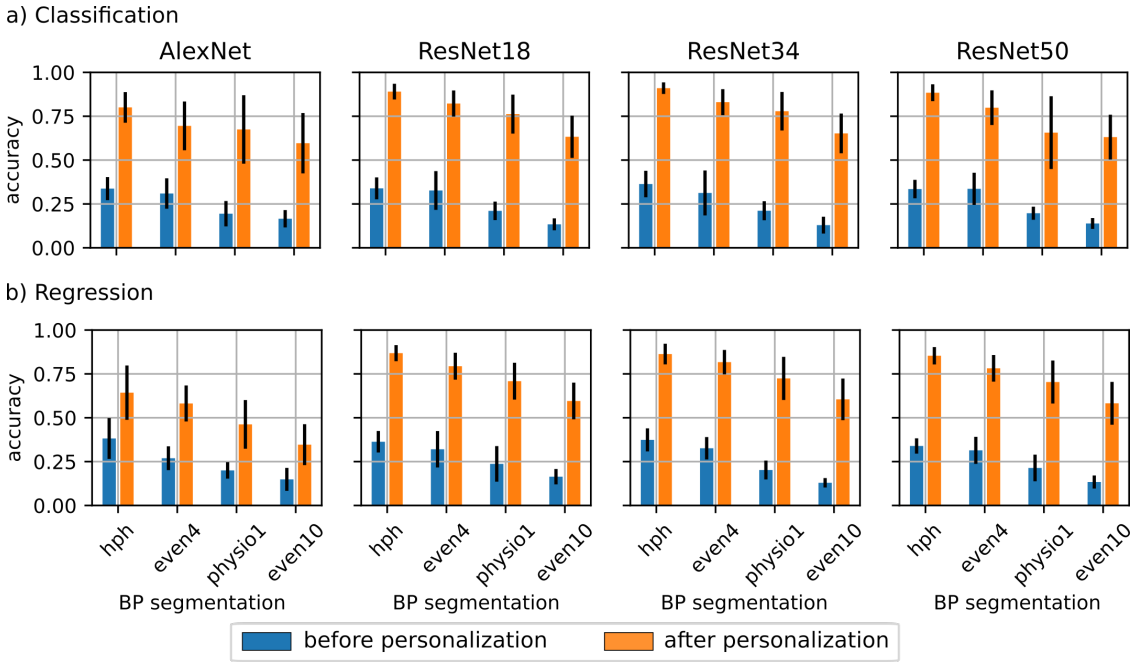

Figure 4. Mean and standard deviation of th test accuracy before (blue) and after (orange) personalization using subject-specific data of 10 test subjects for all considered BP segmentations and every neural architecture. Presented results are divided into results from the classification (a) and regression based (b) approaches.

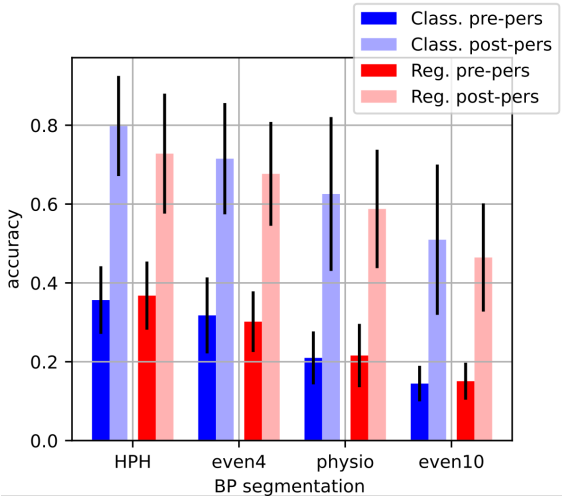

Figure 5. Mean and standard deviation of the test accuracy before and after personalization for all considered BP segmentations. Presented accuracies are averaged over all neural architectures.

post-personalization accuracy was higher when predicting BP bins directly instead of performing a regression and assigning the appropriate BP bin to every predicted SBP-value. This effect seems to be mainly driven by the AlexNet results which show the most pronounced difference in post-personalization test accuracy between the regression and classification based scenario. However, a consistent minor effect can also be seen for all other network architectures. As stated before, test accuracy decreased with increasing numbers of BP bins.

5. Discussion

This paper aimed at a comprehensive comparison of regression-based and classification-based non-invasive BP prediction using deep learning methods. We did not intend to derive a particularly accurate model to achieve state-of-the-art performance. Instead, we adopted a pragmatic approach to explore how the problem of PPG-based BP-prediction might be reformulated to answer relevant questions and be useful in a clinical setting. We derived an extensive dataset from the MIMIC-III database and trained well-established neural network architectures for both BP regression and classification. We divided the SBP into bins of varying number and width (BP range segmentations). The width and number of bins in these segmentations were designed to both cover physiologically relevant BP intervals (e.g. hypo-, normo- and hypertension) as well as to mimic a more regression-like approach (e.g. narrow bins of constant width). NN were trained using loss functions for regression and classification, while the evaluations were carried out with respect to classification scenarios.

Given this experimental setup, our objective was to answer several questions. We investigated whether classification or regression based approaches should be preferred one

over another for BP class prediction in certain scenarios. More precisely, it might be plausible to prefer a classification loss during training for a coarse BP range segmentation and a regression loss in case of a denser segmentation. As our results indicate, this is not the case in general. It can be seen that the performance of both approaches constantly drops from coarser towards denser segmentations. This is obvious from the regression loss perspective since the predicted target variable (continuous BP values) is independent from the particular BP range segmentation. Therefore, the expected mean absolute error is the same for all segmentations which results in a higher chance of misclassifying the target towards denser segmentations, thus leading to a decrease of the performance.

Regarding the pre-personalization results there is no reason to prefer any of the two approaches for any of the given BP range segmentations, although regression methods seem to perform slightly better. This is different for post-personalization results where the classification based methods clearly outperform their regression counterparts. However, there is no indication that classifications based approaches might be less suited for denser BP segmentations compared to regression approaches. Given all of these results, there is also no obvious tradeoff between the coarseness of the BP range segmentation and the use of either regression or classification losses for training.

Our results also emphasize the importance of personalization, i.e. fine tuning the network weights with subject specific data. This personalization procedure, which has been proven effective in various studies before [19, 20, 29], leads to a strong performance increase for any network architectures and their training with both classification and regression losses compared to pre-personalization results. From the results in figure 4 it can be seen that the ResNet architecture benefits the most from personalization. This may be due to the skip connections incorporated in the architecture that allow the network to converge to a better optimum compared to the AlexNet architecture. It seems to be even slightly more effective for classification based approaches since they outperform regression approaches after personalization which is not consistently the case before personalization. However, we could not find indications that the effectiveness of personalization depends on the particular task, i.e. the BP range segmentation.

We acknowledge that the subject population we used for personalization may not be sufficient to draw general conclusions. Additional selection criteria have to be employed to ensure that the subjects are representative for a population spanning a greater range of demographic and medical characteristics. Furthermore, it has to be investigated which properties and patterns enable the NN to improve its classification accuracy after training on subject-specific data. However, our results suggest that classification instead of regression has the potential to greatly improve the accuracy of non-invasive BP prediction.

Classification and regression of BP is subject to ongoing research and several relevant studies reported results that partly fulfilled the criteria of the British Hypertension Society and the Association of the Advancement of Medical Instrumentation. However, special attention has to be paid to the experimental setup and the design of the training pipeline in order to obtain unbiased results that allow for a realistic assessment of the method's clinical applicability. In the light of these considerations, many authors question the practical feasibility of a truly continuous BP estimation [27]. One of the reasons may be external factors (e.g. age, chronical illnesses, medication and differences in measurement equipment) that introduce inter-subject variations into the PPG morphology which prevents the neural networks from good generalization. Given our experimental setup which ensured a balanced number of samples across BP bins in each segmentation, our post-personalization results indicate that it can be possible to achieve a BP classification performance of practical relevance solely based on PPG signals for some application scenarios.

Our results provide three insights: (i) training neural networks from scratch does not lead to an advantage in terms of test accuracy for either classification or regression approaches. Test accuracy drops as the number of BP segments gets larger. (ii) Personalization is immensely important for BP prediction as it enables machine learning methods to identify informative patterns in the subject-specific feature space through training on subject-specific data; (iii) classification-based approaches may be preferable over regression-based approaches when the correct association of BP to a small number of broad BP intervals is sufficient.

The findings in this work are of great importance when applying machine learning methods to camera-based PPG measurements. Such methods experience great interest from the scientific community and are investigated extensively [30, 34, 40]. However, due to confounding factors like movement, reflections or changing lighting conditions that negatively impact the signal-to-noise ratio compared to their sensor based counterpart, it can be expected that the BP prediction error increases. Therefore, developing a method that satisfies all the relevant clinical criteria is still a work in progress. Simplifying the problem by classifying a limited number of BP ranges might be a way forward to arrive at a truly non-invasive BP prediction method that is applicable in a clinical setting.

References

[1] *Guideline for the pharmacological treatment of hypertension in adults*. WHO Guidelines Approved by the Guidelines Re-

view Committee. World Health Organization, Geneva, 2021. 4

[2] Clementine Aguet, Jerome Van Zaen, Joao Jorge, Martin Proenca, Guillaume Bonnier, Pascal Frossard, and Mathieu Lemay. Feature Learning for Blood Pressure Estimation from Photoplethysmography. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 463–466, Mexico, Nov. 2021. IEEE. 3

[3] Sanghyun Baek, Jiyong Jang, and Sungroh Yoon. End-to-End Blood Pressure Prediction via Fully Convolutional Networks. *IEEE Access*, 7:185458–185468, 2019. 1, 2

[4] Jesus Cano, Lorenzo Facila, Philip Langley, Roberto Zangroniz, Raul Alcaraz, and Jose J. Rieta. Application of Deep Neural Network Models for Blood Pressure Classification based on Photoplethysmograpic Recordings. In *2021 International Conference on e-Health and Bioengineering (EHB)*, pages 1–4, Iasi, Romania, Nov. 2021. IEEE. 1, 2, 3

[5] Xiao-Rong Ding, Ni Zhao, Guang-Zhong Yang, Roderic I. Pettigrew, Benny Lo, Fen Miao, Ye Li, Jing Liu, and Yuan-Ting Zhang. Continuous Blood Pressure Measurement From Invasive to Unobtrusive: Celebration of 200th Birth Anniversary of Carl Ludwig. *IEEE Journal of Biomedical and Health Informatics*, 20(6):1455–1465, Nov. 2016. 1

[6] Chadi El Hajj and Panayiotis 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)*, pages 4269–4272, Montreal, QC, Canada, July 2020. IEEE. 1, 3

[7] Heiko Gesche, Detlef Grosskurth, Gert Küchler, and Andreas Patzak. Continuous blood pressure measurement by using the pulse transit time: Comparison to a cuff-based method. *European Journal of Applied Physiology*, 112(1):309–315, Jan. 2012. Publisher: Springer ISBN: 1439-6327 (Electronic)\r1439-6319 (Linking). 3

[8] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. page 8. 5

[9] Ary L Goldberger, Luis A N Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, C.-K. Peng, and H Eugene Stanley. PhysioBank, PhysioToolkit, and PhysioNet : Components of a New Research Resource for Complex Physiologic Signals. *Circulation*, 101(23):e215–e220, June 2000. 3

[10] Serj Haddad, Assim Boukhayma, and Antonino Caizzone. Continuous PPG-Based Blood Pressure Monitoring Using Multi-Linear Regression. *arXiv:2011.02231 [physics]*, Nov. 2020. arXiv: 2011.02231. 1

[11] Latifa Nabila Harfiya, Ching-Chun Chang, and Yung-Hui Li. Continuous Blood Pressure Estimation Using Exclusively Photopletysmography by LSTM-Based Signal-to-Signal Translation. *Sensors*, 21(9):2952, Apr. 2021. 3

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, Las Vegas, NV, USA, June 2016. IEEE. 4

[13] Jingyuan Hong, Jiasheng Gao, Qing Liu, Yuanting Zhang, and Yali Zheng. Deep Learning Model with Individualized Fine-tuning for Dynamic and Beat-to-Beat Blood Pressure Estimation. In *2021 IEEE 17th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*, pages 1–4, Athens, Greece, July 2021. IEEE. 2

[14] Da Un Jeong and Ki Moo Lim. Combined Deep CNN–LSTM Network-based Multitasking Learning Architecture for Noninvasive Continuous Blood Pressure Estimation using Difference in ECG-PPG Features. preprint, In Review, Jan. 2021. 1, 2, 3

[15] Alistair Johnson, Tom Pollard, and Roger Mark. MIMIC-III Clinical Database, 2015. Type: dataset. 3

[16] Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1):160035, Dec. 2016. 3

[17] Koshiro Kido, Zheng Chen, Ming Huang, Toshiyo Tamura, Wei Chen, Naoaki Ono, Masachika Takeuchi, Md. Altaf-Ul-Amin, and Shigehiko Kanaya. Discussion of Cuffless Blood Pressure Prediction Using Plethysmograph Based on a Longitudinal Experiment: Is the Individual Model Necessary? *Life*, 12(1):11, Dec. 2021. 2

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, May 2017. 4

[19] Dongseok Lee, Hyunbin Kwon, Dongyeon Son, Heesang Eom, Cheolsoo Park, Yonggyu Lim, Chulhun Seo, and Kwangsuk Park. Beat-to-Beat Continuous Blood Pressure Estimation Using Bidirectional Long Short-Term Memory Network. *Sensors*, 21(1):96, Dec. 2020. 8

[20] Jared Johann Leitner, Po-Han Chiang, and Sujit Dey. Personalized Blood Pressure Estimation Using Photoplethysmography: A Transfer Learning Approach. *IEEE Journal of Biomedical and Health Informatics*, pages 1–1, 2021. 1, 2, 3, 5, 8

[21] Yongbo Liang, Zhencheng Chen, Rabab Ward, and Mohamed Elgendi. Photoplethysmography and Deep Learning: Enhancing Hypertension Risk Stratification. *Biosensors*, 8(4):101, Oct. 2018. 3

[22] Sakib Mahmud, Nabil Ibtehaz, Amith Khandakar, Anas M. Tahir, Tawsifur Rahman, Khandaker Reajul Islam, Md Shafayet Hossain, M. Sohel Rahman, Farayi Musharavati, Mohamed Arselene Ayari, Mohammad Tariqul Islam, and Muhammad E. H. Chowdhury. A Shallow U-Net Architecture for Reliably Predicting Blood Pressure (BP) from Photoplethysmogram (PPG) and Electrocardiogram (ECG) Signals. *Sensors*, 22(3):919, Jan. 2022. 1, 3

[23] Sumbal Maqsood, Shuxiang Xu, Son Tran, Saurabh Garg, Matthew Springer, Mohan Karunanithi, and Rami Mohawesh. A survey: From shallow to deep machine learning approaches for blood pressure estimation using biosensors. *Expert Systems with Applications*, 197:116788, July 2022. 1

[24] Elisa Mejía-Mejía, James M. May, Panayiotis A. Kyriacou, and Mohamed Elgendi. Classification of blood pressure in critically ill patients using photoplethysmography and machine learning. *Computer Methods and Programs in Biomedicine*, 208:106222, Sept. 2021. 3

[25] Annunziata Paviglianiti, Vincenzo Randazzo, Giansalvo Cirrincione, and Eros Pasero. Neural Recurrent Approches to Noninvasive Blood Pressure Estimation. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7, Glasgow, United Kingdom, July 2020. IEEE. 1

[26] Solmaz Rastegar, Hamid GholamHosseini, Andrew Lowe, and Maria Lindén. A Novel Convolutional Neural Network for Continuous Blood Pressure Estimation. In Tomaz Jarm, Aleksandra Cvetkoska, Samo Mahnič-Kalamiza, and Damijan Miklavcic, editors, *8th European Medical and Biological Engineering Conference*, volume 80, pages 22–28, Cham, 2021. Springer International Publishing. Series Title: IFMBE Proceedings. 3

[27] Fabian Schrumpf, Patrick Frenzel, Christoph Aust, Georg Osterhoff, and Mirco Fuchs. Assessment of Non-Invasive Blood Pressure Prediction from PPG and rPPG Signals Using Deep Learning. *Sensors*, 21(18):6022, Sept. 2021. 2, 3, 8

[28] Umit Senturk, Kemal Polat, and Ibrahim Yucedag. A non-invasive continuous cuffless blood pressure estimation using dynamic Recurrent Neural Networks. *Applied Acoustics*, 170:107534, Dec. 2020. 3

[29] Gašper Slapničar, Nejc Mlakar, and Mitja Luštrek. Blood Pressure Estimation from Photoplethysmogram Using a Spectro-Temporal Deep Neural Network. *Sensors*, 19(15):3420, Aug. 2019. 1, 2, 3, 5, 8

[30] Pedro Henrique de Brito Souza, Israel Machado Brito Souza, Symone Gomes Soares Alcalá, Priscila Valverde de Oliveira Vitorino, Adson Ferreira da Rocha, and Talles Marcelo Gonçalves de Andrade Barbosa. Video-based Photoplethysmography and Machine Learning Algorithms to Achieve Pulse Wave Velocity. *International Journal of Biotech Trends and Technology*, 11(1):7–15, Feb. 2021. 1, 8

[31] Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. Highway Networks. 2015. Publisher: arXiv Version Number: 2. 4

[32] Timo E Strandberg and Kaisu Pitkala. What is the most important component of blood pressure: systolic, diastolic or pulse pressure? *Current opinion in nephrology and hypertension*, 12(3):293–297, 2003. 4

[33] Xiaoxiao Sun, Liang Zhou, Shendong Chang, and Zhaohui Liu. Using CNN and HHT to Predict Blood Pressure Level Based on Photoplethysmography and Its Derivatives. *Biosensors*, 11(4):120, Apr. 2021. 1, 2, 3

[34] Ryo Takahashi, Keiko Ogawa-Ochiai, and Norimichi Tsumura. Non-contact method of blood pressure estimation using only facial video. *Artificial Life and Robotics*, July 2020. 1, 8

[35] Md. Sayed Tanveer and Md. Kamrul Hasan. Cuffless blood pressure estimation from electrocardiogram and photoplethysmogram using waveform based ANN-LSTM network. *Biomedical Signal Processing and Control*, 51:382–392, May 2019. 1

[36] Bryan Williams, Giuseppe Mancia, Wilko Spiering, Enrico Agabiti Rosei, Michel Azizi, Michel Burnier, Denis L Clement, Antonio Coca, Giovanni de Simone, Anna Dominiczak, Thomas Kahan, Felix Mahfoud, Josep Redon, Luis Ruilope, Alberto Zanchetti, Mary Kerins, Sverre E Kjeldsen, Reinhold Kreutz, Stephane Laurent, Gregory Y H Lip, Richard McManus, Krzysztof Narkiewicz, Frank Ruschitzka, Roland E Schmieder, Evgeny Shlyakhto, Costas Tsioufis, Victor Aboyans, Ileana Desormais, and ESC Scientific Document Group. 2018 ESC/ESH Guidelines for the management of arterial hypertension. *European Heart Journal*, 39(33):3021–3104, Sept. 2018. 4

[37] Carl F. Wippermann, Dietmar Schranz, and Ralph 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*, 11(5):324–328, Sept. 1995. 3

[38] Jiaze Wu, Hao Liang, Changsong Ding, Xindi Huang, Jianhua Huang, and Qinghua Peng. Improving the Accuracy in Classification of Blood Pressure from Photoplethysmography Using Continuous Wavelet Transform and Deep Learning. *International Journal of Hypertension*, 2021:1–9, Aug. 2021. 1, 3

[39] Guanqun Zhang, Mingwu Gao, Da Xu, N. Bari Olivier, and Ramakrishna Mukkamala. Pulse arrival time is not an adequate surrogate for pulse transit time as a marker of blood pressure. *Journal of Applied Physiology*, 111(6):1681–1686, Dec. 2011. 3

[40] Jiancheng Zou, Shouyu Zhou, Bailin Ge, and Xin Yang. Non-Contact Blood Pressure Measurement Based on IPPG. *Journal of New Media*, 3(2):41–51, 2021. 1, 8