Personalized Blood Pressure Estimation Using Photoplethysmography: A Transfer Learning Approach

Jared Leitner, Po-Han Chiang, *Student Member, IEEE*, and Sujit Dey, *Fellow, IEEE*

***Abstract*—In this paper, we present a personalized deep learning approach to estimate blood pressure (BP) using the photoplethysmogram (PPG) signal. We propose a hybrid neural network architecture consisting of convolutional, recurrent, and fully connected layers that operates directly on the raw PPG time series and provides BP estimation every 5 seconds. To address the problem of limited personal PPG and BP data for individuals, we propose a transfer learning technique that personalizes specific layers of a network pre-trained with abundant data from other patients. We use the MIMIC III database which contains PPG and continuous BP data measured invasively via an arterial catheter to develop and analyze our approach. Our transfer learning technique, namely BP-CRNN-Transfer, achieves a mean absolute error (MAE) of 3.52 and 2.20 mmHg for SBP and DBP estimation, respectively, outperforming existing methods. Our approach satisfies both the BHS and AAMI blood pressure measurement standards for SBP and DBP. Moreover, our results demonstrate that as little as 50 data samples per person are required to train accurate personalized models. We carry out Bland-Altman and correlation analysis to compare our method to the invasive arterial catheter, which is the gold-standard BP measurement method.**

***Index Terms*—Deep learning, transfer learning, wearables, blood pressure, photoplethysmogram.**

I. Introduction

BLOOD pressure (BP) is the most important indicator of cardiovascular health. High blood pressure, or hypertension, affects 30% of American adults and contributes to over 410,000 deaths per year [1], [2]. This condition has been called “the silent killer,” as typically no symptoms are recognized before significant damage has already been done to the heart and arteries [3]. BP is defined as the pressure exerted on the arteries as blood is pumped throughout the body and is measured in millimeters of mercury (mmHg). Systolic (SBP) and diastolic blood pressure (DBP) are the primary metrics used to measure BP, which are defined as the maximum and minimum blood pressure, respectively, during a pulse.

For accurate diagnosis and treatment of hypertension, regular BP measurement is necessary. According to the American College of Cardiology, increased at-home BP monitoring is essential for recognizing inconsistencies in measurements taken in a medical setting [4]. Currently, the predominant device for measuring BP is a mercury sphygmomanometer which involves attaching an inflatable cuff around the upper arm [5]. This process requires significant user effort, which limits the frequency of BP measurements and increases the chance of measurement error. The use of an arterial catheter can provide continuous BP measurement; however, it is highly invasive and impractical for daily life. On the other hand, wearable devices are widely used for non-invasive, continuous monitoring of biological information [6]. Continuous and automated blood pressure estimation could be incorporated into one’s daily routine to obtain better insight and detect abnormal BP fluctuation.

One prominent approach is to estimate BP with the photoplethysmogram (PPG) sensor, which is available in most wrist wearables. The principle of the PPG sensor is to optically measure the dilation and constriction of blood vessels. The resulting PPG signal is a fusion of heart activity, vascular relaxation processes, and microcirculation system status, making its time-frequency domain information rich and diverse [7]. In this paper, we propose a deep learning approach to personalized BP estimation based on the PPG signal.

A. Related Work

Traditional machine learning approaches to PPG-based BP estimation focus on pulse wave analysis (PWA) methods. PWA involves extracting both time and frequency domain features from the PPG series and using these hand-crafted features as inputs to the BP estimation model. [8] extracts nineteen features from each PPG cycle based on its morphology. They use these features and the corresponding SBP and DBP values to train different regression models. Their approach lacks personalization, which may be the reason for higher estimation errors since these features have a person-specific response to BP [9]. [10] and [11] both use a random forest as their BP estimation model. [10] uses a feature selection algorithm to determine which morphological features are most useful for BP estimation and found that many

Manuscript received November 27, 2020; revised March 26, 2021 and May 19, 2021; accepted May 25, 2021. Date of publication June 2, 2021; date of current version January 5, 2022. This work was supported in part by the UC San Diego Center for Wireless Communications, Teradata, and in part by Samsung Research America. *(Corresponding author: Jared Leitner.)*

The authors are with the Electrical and Computer Engineering Department, Jacobs School of Engineering, University of California, San Diego, CA 92093 USA (e-mail: jjleitne@eng.ucsd.edu; pochiang@ucsd.edu; dey@ece.ucsd.edu).

Digital Object Identifier 10.1109/JBHI.2021.3085526

features are irrelevant. Since the PPG signal is highly sensitive to different sources of noise [12] and its morphology can range from person to person, it is difficult to detect the key points in the signal required for feature engineering. In addition, manually engineered features can prove to be redundant or irrelevant in the PPG-BP modeling process. As a result, the information contained in the PPG signal may not be fully utilized.

In our previous work [13], we propose a method for personalized BP estimation using wavelet decomposition to extract time-frequency domain features from the PPG signal. These features are then used to train a random forest model for SBP and DBP estimation. Unlike previous approaches which extract features from the PPG signal on a per cycle basis, wavelet decomposition captures dependencies between cycles in the time-frequency domain. While this approach produced accurate estimations, 10 hours of continuous BP and PPG data are required per person for training. Although PPG data can be continuously measured, large amounts of BP data are difficult to acquire outside a hospital setting.

In order to address the limitations of these previous methods, we propose a deep learning approach that utilizes a novel transfer learning technique that requires as little as 50 samples to train accurate personalized models. Deep learning models are widely used to model nonlinear relationships and have been applied to various tasks involving physiological signals [14]–[16]. Deep learning addresses the challenges of manual feature engineering and information loss by directly learning from the raw PPG data. [17]–[19] build deep learning models for PPG-based BP estimation and utilize personalization techniques to improve performance. [17] uses a spectro-temporal neural network that takes a 5 second PPG segment and its corresponding spectrogram as inputs to their model. When personalizing their model, the SBP and DBP MAE decrease by 39% and 44%, respectively, indicating that the relationship between BP and PPG is subject-dependent. [18] utilizes a Siamese neural network to estimate the offset from a calibration PPG-BP sample. The network uses a series of convolutional layers to derive an effective representation of the PPG series and achieves high estimation performance. [19] proposes a convolutional neural network (CNN) for BP estimation and utilizes transfer learning to personalize their model to each patient. Their proposed model requires 4000 personal BP samples for transfer learning to achieve high performance. Such a large number of personal BP samples is not possible to collect outside a hospital setting.

Transfer learning focuses on storing knowledge gained from solving one problem (i.e., source domain) and applying it to a different but related problem (i.e., target domain), which usually contains a small number of data samples to train a model [20]. We propose to use a pre-trained model with abundant PPG and BP data from a large pool of source patients to drastically reduce the required data for new patients, as illustrated in Fig. 1.

Deep learning models are conducive to transfer learning due to the modularity of their architectures. In this work, we develop our architecture, namely Blood Pressure – Convolutional Recurrent Neural Network (BP-CRNN), based on the Convolutional, Long Short-Term Memory, fully connected Deep Neural Network (CLDNN) [21], one of the popular hybrid artificial neural network (ANN) architectures. Our proposed method, namely BP-CRNN-Transfer, personalizes specific network layers during transfer learning to reduce the number of required training samples. Our contributions are as follows:

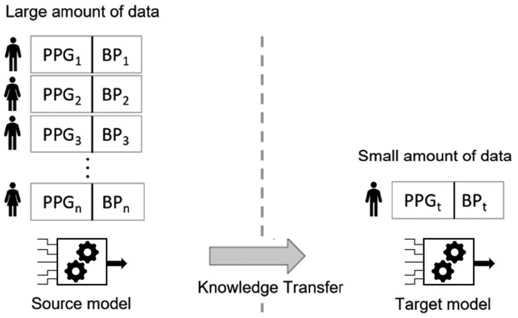


Fig. 1. Transfer learning overview for PPG-based BP estimation.

- We propose a hybrid neural network consisting of convolutional and recurrent layers which operate directly on the raw PPG time series to reduce information loss and effectively model the PPG-BP relationship.
- We propose a novel transfer learning technique that personalizes specific layers of a pre-trained network to improve the performance of PPG-based BP estimation, demonstrating that PPG-BP data of other patients can be used to enhance the modeling of a new patient's PPG-BP relationship.
- We demonstrate that the proposed transfer learning technique improves BP estimation performance by 23.3% for SBP and 19.1% for DBP. We verify our approach is consistent with the gold-standard BP measurement method through Bland-Altman and correlation analysis.
- We show that our proposed transfer learning method requires 10x less personal PPG-BP data to achieve performance equivalent to that of a new personalized model trained with abundant data.

The rest of the paper is organized as follows. In Section II, data acquisition and our network architecture are presented. We then detail the proposed transfer learning technique. In Section III, the performance of the proposed method is evaluated. We compare how model performance changes for different numbers of training samples, with and without using transfer learning. Finally, we conclude the paper in Section IV.

II. Method

In this section, we first describe the MIMIC III Matched Subset database and the PPG and BP preprocessing steps. We then present the network architecture and transfer learning technique.

A. Data Acquisition and Preprocessing

Data was obtained from the Multiparameter Intelligent Monitoring in Intensive Care III (MIMIC III) Matched Subset database [22], [23]. This database contains records for thousands of intensive care unit patients. Records in this database have been matched to records from the MIMIC III Clinical database [24],

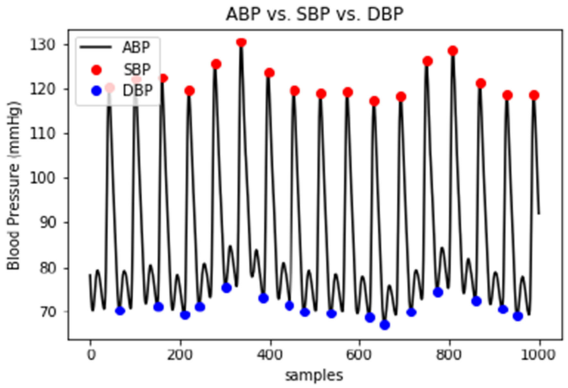


Fig. 2. Output of peak detection algorithm – SBP and DBP vs. raw ABP time series.

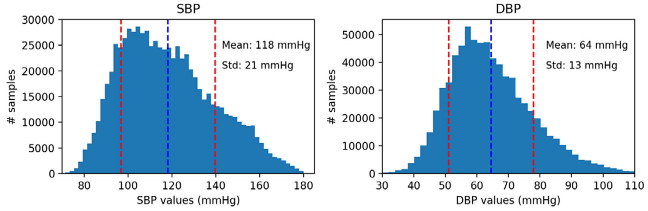


Fig. 3. Distribution of SBP and DBP samples among the 100 patients. The blue dashed lines indicate the mean SBP/DBP and the red dashed lines correspond to 1 standard deviation above and below the mean SBP/DBP.

which includes de-identified demographic data. The waveforms collected include ECG, respiration, continuous blood pressure, and PPG signals each sampled at 125 Hz. The arterial blood pressure (ABP) was directly measured from a radial artery using an invasive catheter. A fingertip sensor was used to measure the PPG data. Only patients with sufficient PPG and blood pressure data were considered for this study. We trained and tested our PPG-based BP estimation method on 100 randomly selected patients who had at least 10 hours of high-quality data after preprocessing. Out of these 100 patients, 56 are male and 44 are female. The age of the patients ranges from 21 to 82 with a mean age of 58.

Our objective is to operate directly on the raw PPG data and estimate SBP and DBP simultaneously. The first stage of data preprocessing involves splitting the raw PPG signal into 5-second segments and down sampling from 125 Hz to 25 Hz as this covers the important frequency components [25]. Next, each PPG segment is labeled with the mean SBP and DBP during that segment. SBP and DBP values are obtained from the raw ABP series using a peak detection algorithm as illustrated in Figure 2. Figure 3 describes the distribution of SBP and DBP samples. Some sections of the PPG series are corrupted due to motion artefacts or because the patient was not properly wearing the sensor. In order to discard these corrupted sections, an autocorrelation filter is implemented. Since an uncorrupted PPG segment should maintain a high degree of periodicity, it is expected that the signal's autocorrelation is high when the segment is offset by multiples of the cycle length. Figure 4 displays both an uncorrupted and corrupted PPG segment and the corresponding autocorrelation signals. The peaks in the autocorrelation signal are used to determine the quality of each PPG segment. An empirical threshold of 0.7 was set on the maximum autocorrelation. The filtered PPG segments are then normalized to zero mean and unit variance. Using this labeled dataset, we train our proposed personalized deep neural networks for BP estimation.

B. Network Architecture

We propose a hybrid network architecture, namely BP-CRNN, that makes use of convolutional layers, a gated recurrent unit (GRU), and fully connected (FC) layers. This is an adaptation of the CLDNN network presented in [21]. Instead of a LSTM, we use a GRU which behaves nearly identically with one fewer equation. In addition, we pass the outputs of both the first and third convolutional layers to the GRU. Figure 5 displays our architecture. The rationale is as follows: The convolutional layers serve as feature extractors for the raw PPG input, while the GRU models the temporal dependencies between these features. The GRU's outputs are then fed to the fully connected layers which transform the features into a space that makes the BP easier to estimate.

The input PPG segment is convolved with 50 different filters to generate 50 outputs in the temporal-feature domain. The following two convolutional layers also contain 50 filters, which are convolved with these features to generate the final features from the PPG segment. Each layer is followed by a rectified linear unit (ReLU) activation function. The output feature maps of each convolutional layer are calculated using the equation:

$$x_j^l = Relu\left(\left(\sum_i x_i^{l-1} * k_{ij}\right) + b_j^i\right) \tag{1}$$

where x_j^l is the j_{th} map generated by the convolutional layer l, x_i^{l-1} is the i_{th} feature map of the previous convolutional layer $l-1$, k_{ij} represents the i_{th} trained convolution kernel, b_j^i is the additive bias, while $*$ represents the convolution operation and *Relu* is the activation function.

Stacking convolutional layers results in a learned feature hierarchy, where initial layers extract lower-level features and deeper layers extract higher-level features [26]. We varied the number of convolutional layers from 1 to 5 and found that 3 convolutional layers resulted in the best performance. In order to provide both low and high-level features to the GRU to process simultaneously, the outputs of the first and third convolutional layers are concatenated. Since each convolutional layer contains 50 filters, 100 extracted feature series are passed to the GRU. The extracted features at each level are padded such that they have the same length as the input PPG sequence. As a result, the input to the GRU has a shape of $100 * t_n$ where t_n is the length of the input PPG segment. The GRU is able to learn the temporal relationship between these multiple feature channels. A GRU consists of gating units that control the flow of information

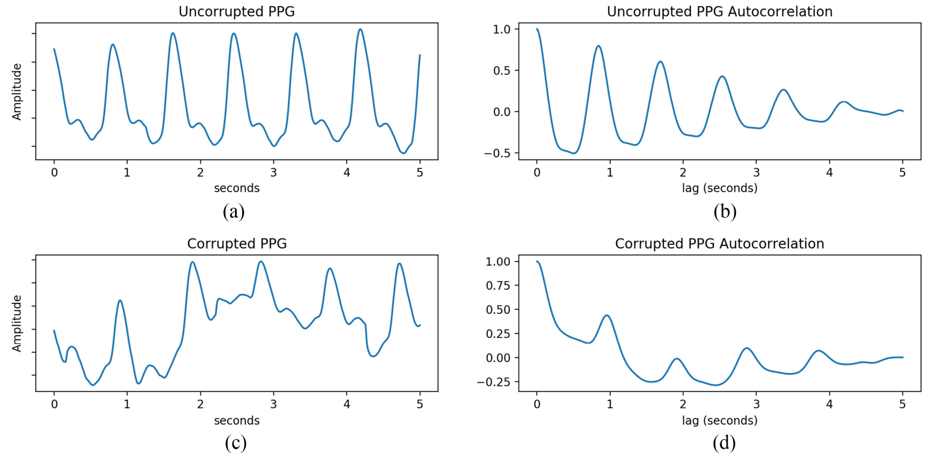


Fig. 4. Comparison of (a) an uncorrupted PPG segment and (b) its corresponding autocorrelation signal to (c) a corrupted PPG segment and (d) its corresponding autocorrelation signal.

within the module [27]. The following equations describe the operation of the GRU:

$$z_t = \sigma\left(W^{(z)}x_t + U^{(z)}h_{t-1}\right) \tag{2}$$

$$r_t = \sigma\left(W^{(r)}x_t + U^{(r)}h_{t-1}\right) \tag{3}$$

$$h'_t = tanh\left(W^{(h)}x_t + U^{(h)}(r_t \odot h_{t-1})\right) \tag{4}$$

$$h_t = z_t \odot h_{t-1} + (1 - z_t) \odot h'_t \tag{5}$$

In (5), the final GRU activation h_t is a linear interpolation between the previous activation h_{t-1} and candidate activation h'_t where the update gate z_t determines how much the unit updates its activation. $\odot$ represents element-wise multiplication. Eq. (2) describes the update gate z_t calculation, where $W^{(z)}$ and $U^{(z)}$ are each a set of trainable weights that process the input x_t and the previous activation h_{t-1}, respectively. σ represents the sigmoid function. The candidate activation h'_t is calculated in Eq. (4), where r_t represents the reset gate, $W^{(h)}$ and $U^{(h)}$ represent trainable sets of weights, and $tanh$ represents the hyperbolic tangent function. When r_t is close to 0, the reset gate enables the unit to forget the previous activation h_{t-1} when calculating the candidate activation h'_t [27]. In Eq. (3), the reset gate r_t is calculated similarly to the update gate. $W^{(r)}$ and $U^{(r)}$ represent the reset gate's trainable weights that process the input x_t and the previous activation h_{t-1}, respectively. At each time step, a 100-element vector is processed by the GRU, where each element corresponds to a feature value. A GRU activation size of 25 was experimentally determined to produce high performance, resulting in an output of shape $25 * t_n$.

The last two network layers are fully connected layers that carry out the final BP estimation. FC layers are effective at mapping features into a more separable space [26]. The activations of the GRU at each time step are flattened into a single vector for the first FC layer to the process. The output of the network is a 2-dimensional vector corresponding to the estimated SBP and DBP values. A ReLU activation function is again used after each FC layer. Batch normalization [28] is utilized to stabilize the input distribution of each layer during training. This reduces internal covariate shifts and results in faster training. Overall, this architecture realizes the high level of complementarity these individual neural network layers exhibit.

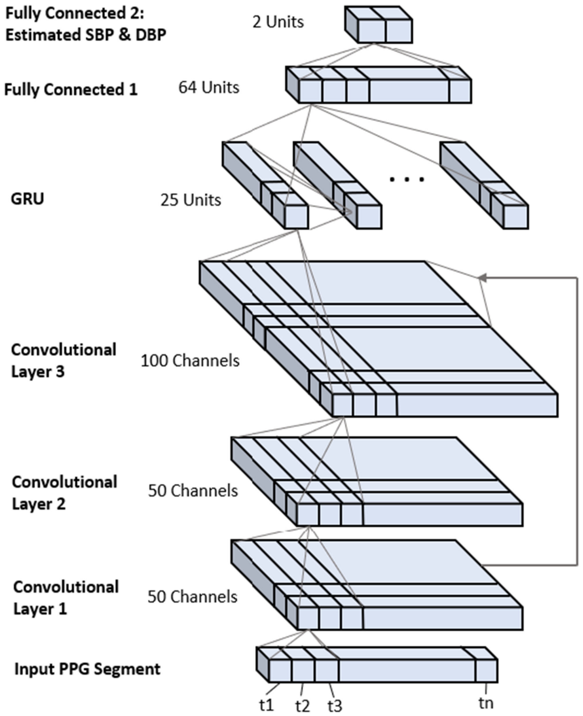


Fig. 5. Proposed BP-CRNN architecture– Convolutional layers serve as feature extractors, GRU models temporal relationship between features, and fully connected layers transform GRU outputs to SBP and DBP.

C. Transfer Learning

To train deep neural networks, a large amount of training data is required to learn effective feature representations. Since our goal is to train personalized PPG-based BP estimation models, this means many data samples from a single individual are

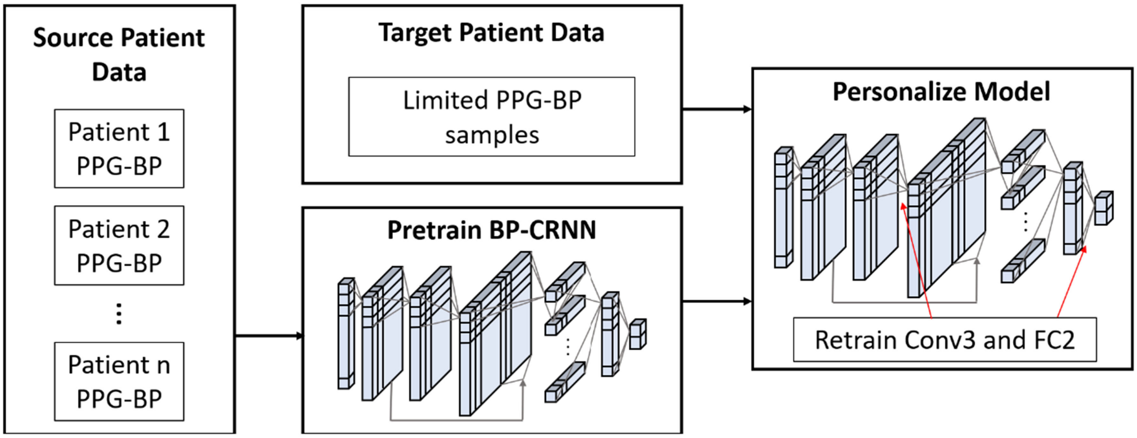


Fig. 6. Proposed transfer learning method, namely BP-CRNN-Transfer. A BP-CRNN model is first pretrained using abundant source patient data. The final convolutional layer and fully connected layer are finetuned with the target patient's data.

required. While PPG data can continuously be collected via a noninvasive wearable, BP data is more difficult to collect. In order to address this, we propose a transfer learning technique that results in high performance even when limited data from the target patient is available.

Transfer learning has most notably been applied to computer vision (CV) and natural language processing (NLP) tasks. [29] argues that physiological signals share two important commonalities with CV and NLP: consistency and complexity. Physiological patterns are consistent across individuals but complex enough that learning effective feature representations is nontrivial. [30] describes how initial convolutional layers extract lower-level features, which can be shared across tasks, while deeper layers generate higher-level features which are task-specific. In addition, training with different tasks (patients in our case), can result in a more powerful representation of the data that could not be learned from a single task (patient). Inspired by [29], [30], we first train our model with PPG-BP data from a variety of individuals to learn robust feature extractors that can be transferred between patients.

Figure 6 illustrates our proposed transfer learning process, namely BP-CRNN-Transfer. PPG and BP data from n source patients is used to pre-train a BP-CRNN model. This network is then used as an initialization for finetuning. In order to personalize the model, data from the target patient is used to finetune specific layers in the network. The last convolutional layer (Conv3) and last fully connected layer (FC2) are retrained using the target patient's data. In addition, the batch normalization parameters are updated to account for the different data distribution of the target patient. It was experimentally determined that retraining these two specific layers resulted in the most robust transfer learning performance. Table III in Sec. III (B) describes the transfer learning performance for different combinations of personalized layers. By retraining the final convolutional layer, the network can learn high-level PPG feature representations specific to the individual. Finetuning the last FC layer allows the model to learn the relationship between the extracted features and BP for the patient of interest. Our BP-CRNN model consists of approximately 250,000 trainable parameters, where 18,000 of these parameters are within the two layers we finetune. This indicates that we only need to update 7.2% of the network parameters learned from the source dataset. By retraining a small percentage of parameters, we prevent the network from overfitting to the limited target training data.

III. Results and Discussion

In this section, we describe the experiment settings and compare our personalized BP estimation results with and without transfer learning to previous methods. We examine how performance is affected by the number of personal data samples used during training and demonstrate that our transfer learning approach can achieve high performance with limited data. We verify our approach is consistent with the gold-standard BP measurement method through Bland-Altman and correlation analysis.

A. Experiment Setting

We implement and evaluate our deep learning model using the Pytorch library [31] in the python environment on an Intel i5 3.2GHz quad-core and 16GB RAM computer. Nvidia GeForce GPUs are utilized to carry out network training. 1-dimensional filters of size 7 were implemented for each convolutional layer and zero padding was used to maintain the input PPG dimension. Based on the results from [32], a large range in the number of filters will result in similar performance before overfitting occurs. We chose to use 50 filters at each layer. All networks are trained using the Adam optimizer [33]. 10 hours of PPG and BP data are selected from each patient to be used in our experiments. 5-fold cross-validation is carried out for each patient separately. This involves shuffling each patient's data and using 5 different train, validation, and test splits for each experiment. Each validation and test set comprises of 1 hour of PPG-BP data. The number of samples included in the training sets is varied from 50 to 3600 samples in order to assess how performance is affected by training set size, which is detailed in Sec. III (C). Data separation between patients is maintained to ensure that no personal data from the target patient is used in pretraining for transfer learning. Mean absolute error (MAE) is calculated and used as our evaluation metric. For each experiment, we provide the average of MAEs over all patients. MAE is defined as follows:

$$MAE = \frac{\sum_{i=1}^{n} \left| BP_{pred}^{i} - BP_{actual}^{i} \right|}{n} \tag{6}$$

For our non-transfer method, namely BP-CRNN, separate personalized models are trained for each of the 100 patients. Each model is trained only using data from the individual patient. Since we do not use transfer learning, the parameters of the initial model are randomly initialized and all layers are updated during training. To train these models, we use 0.01 as the learning rate and 32 as the batch size.

For testing our transfer learning technique, namely BP-CRNN-Transfer, the initial model for the first 50 patients is trained with the data of the last 50 patients, and vice versa. This ensures that no data from the target patient is used during pretraining. When training the initial model for transfer learning, the learning rate and batch size are set to 0.001 and 256, respectively. In this case, the learning rate can be decreased and the batch size increased because there is much more training data, resulting in a greater number of update steps per epoch. When fine-tuning the pre-trained model to the target patient, the learning rate and batch size are set back to 0.01 and 32, respectively, and only the specific layers mentioned in Sec. II (C) are updated. Early stopping [34] is implemented for every training session to save the learned network weights once the error on the validation set begins to increase. Each network is trained 5 times and the results averaged in order to account for differences in model convergence. Our model's inference time is 0.32 ± 0.09 (mean ± std) seconds. This time is based on implementation on a Nvidia GPU. In our future work, we plan to investigate a lightweight model that can be directly implemented on a wearable device and research the tradeoffs between model accuracy, inference time, and memory requirements.

B. BP Estimation Results

We compare the BP estimation performance of our personalized models without and with transfer learning to that of an aggregate model and previous methods in Table I. BP-CRNN and BP-CRNN-Transfer correspond to our personalized approach without and with transfer learning, respectively. The aggregate model, namely Aggregate BP-CRNN, is trained in the same fashion as the pre-trained models for transfer learning as described in the previous section. However, no personalization or transfer learning is applied. The high estimation error of Aggregate BP-CRNN demonstrates the requirement for personalization in order to effectively model the PPG-BP relationship.

Next, we compare our proposed approach against a dummy regressor, namely Mean Regressor, which always predicts the mean SBP and DBP from the target patient's training set. This is an important comparison to make as there may be a subject with relatively constant BP, in which case the BP-CRNN's estimation errors will be low [17]. This comparison is drawn to ensure that our model has learned more than simply estimating the patient's mean BP. In addition, we compare our approach to our previous work and to the latest deep learning approaches that propose personalized BP estimation methods. In our previous work, we apply wavelet decomposition to the PPG series for feature engineering and train a random forest (RF) as our BP estimation model [13]. As mentioned in the introduction section, [17] trains a spectro-temporal neural network using personal data samples from each patient. [18] uses a Siamese neural network that takes a raw PPG segment as input and estimates the BP offset from a calibration PPG-BP sample. [19] trains a convolutional neural network for BP estimation and utilizes transfer learning to personalize their model to each patient.

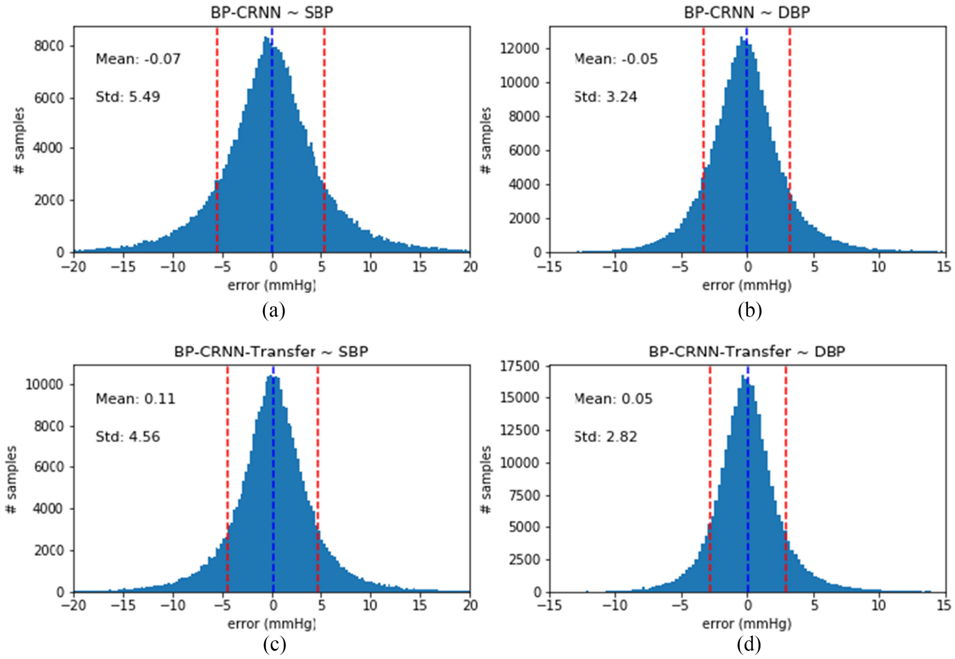
TABLE I
COMPARISON OF BP ESTIMATION METHODS

Method	SBP MAE (mmHg)	DBP MAE (mmHg)
Aggregate BP-CRNN	16.3	8.46
Mean Regressor	9.07	4.58
RF - Wavelet [13]	4.88	2.61
Spectro-Temporal NN [17]	9.43	6.88
Siamese NN [18]	5.95	3.41
CNN-Transfer [19]	4.06	2.20
BP-CRNN	4.59	2.72
BP-CRNN-Transfer	**3.52**	**2.20**

In our current approach, a model is trained for each patient using both a non-transfer learning and transfer learning approach, as described in the experiment setting. Without transfer learning, namely BP-CRNN, we achieve an average MAE of 4.59 and 2.72 mmHg for SBP and DBP, respectively. As shown in Table I, even without using transfer learning, our proposed model achieves improvement in SBP performance compared to the methods presented in [13], [17], [18]. We attribute this improvement to the complementarity of our network architecture and its ability to reduce information loss by operating directly on the raw PPG series. With the transfer learning approach, namely BP-CRNN-Transfer, the MAEs decrease to 3.52 and 2.20 mmHg corresponding to a 23.3% and 19.1% increase in performance for SBP and DBP estimation as compared to our non-transfer method. The performance achieved by our BP-CRNN-Transfer method is also better than our previous approach RF-wavelet [13] as well as previous deep learning methods [17]–[19]. We achieve a 27.9% and 15.7% improvement from [13], 62.7% and 68% improvement from [17], and 40.8% and 35.5% improvement from [18] for SBP and DBP, respectively. We achieve a 13.3% improvement for SBP MAE and the same DBP MAE as compared to [19]. We attribute this increase in performance to the specific layers we finetune during transfer learning and our network's ability to effectively store information contained in source patients' data. The BP-CRNN-Transfer MAE is well under the Mean Regressor MAE, which is 9.07 mmHg for

TABLE II
COMPARISON OF PROPOSED METHOD TO BHS STANDARDS. BOTH OUR NON-TRANSFER (BP-CRNN) AND TRANSFER LEARNING (BP-CRNN-TRANSFER) APPROACHES ACHIEVE GRADE A PERFORMANCE FOR SBP AND DBP

	SBP				DBP			
Method	≤ 5 mmHg	≤ 10 mmHg	≤ 15 mmHg	Grade	≤ 5 mmHg	≤ 10 mmHg	≤ 15 mmHg	Grade
BP-CRNN	72%	92%	97%	A	89%	98%	99%	A
BP-CRNN-Transfer	80%	95%	98%	A	93%	99%	100%	A



(a) (b)

(c) (d)

Fig. 7. Distributions of (a) SBP and (b) DBP errors using our non-transfer approach compared to distributions of (c) SBP and (d) DBP errors using our transfer learning approach. The blue dashed lines indicate the mean error and the red dashed lines correspond to 1 standard deviation above and below the mean error.

SBP and 4.58 mmHg for DBP, indicating that the model can learn a meaningful relationship between PPG and BP. Since [19] achieves the closest performance to our proposed method, we reimplement their approach in order to perform statistical tests. We carry out a Paired Student's *t*-Test separately for each patient to assess the statistical significance of the difference in estimation errors between our method and [19]. For 84 out of the 100 patients, the difference in performance is statistically significant at the level 0.05 for both SBP and DBP.

We evaluate our proposed method according to the British Hypertension Society (BHS) and the Association for the Advancement of Medical Instrumentation (AAMI) standards for BP measurement. The BHS standard assigns a performance grade based on the percentage of estimated BP samples that fall within 5, 10, and 15 mmHg of the corresponding reference BPs. To achieve Grade A accuracy, at least 60/85/95% of the estimated BP samples must have an absolute difference of ≤5/10/15 mmHg from the reference BPs, respectively [35]. Table II describes the results of our non-transfer and transfer learning approaches according to the BHS standards. For our non-transfer approach, 72/92/97% of estimated SBP samples have an absolute difference ≤5/10/15 mmHg, respectively. When using our transfer learning approach, these percentages increase to 80/95/98% of estimated SBP samples. For our non-transfer approach, 89/98/99% of estimated DBP samples have an absolute difference ≤5/10/15 mmHg, respectively. When using our transfer learning approach, these percentages increase to 93/99/100% of estimated DBP samples. Both approaches achieve Grade A performance according to the BHS standard for SBP and DBP.

The AAMI standard for accurate BP measurement requires that the mean error between estimated and reference BPs is ≤5 mmHg and the standard deviation (SD) of errors is ≤8 mmHg [36]. Figure 7 displays the error distribution for SBP and DBP using both our non-transfer and transfer learning approach over all patients. Our BP-CRNN (non-transfer) approach achieves a mean error and standard deviation of -0.07 ± 5.49 mmHg and -0.05 ± 3.24 mmHg for SBP and DBP, respectively. Our BP-CRNN-Transfer approach achieves a mean error and standard deviation of 0.11 ± 4.56 mmHg and 0.05 ± 2.82 mmHg for SBP and DBP, respectively. The mean error for each approach is approximately 0 mmHg. When using our transfer learning approach, the SD of errors decreases from 5.49 to 4.56 mmHg

TABLE III
COMPARISON OF TRANSFER LEARNING PERFORMANCE WHEN FINETUNING DIFFERENT NETWORK LAYERS

BP-CRNN Layers Personalized	SBP MAE (mmHg)	DBP MAE (mmHg)
FC1, FC2	5.16	2.87
FC2	4.41	2.63
Conv1, Conv2, Conv3, GRU, FC1, FC2	4.37	2.41
Conv3, FC1, FC2	4.32	2.46
Conv2, Conv3, GRU, FC1, FC2	4.28	2.38
Conv3, GRU, FC1, FC2	4.25	2.37
Conv3, GRU, FC2	3.90	2.28
Conv3, FC2	**3.84**	**2.24**

TABLE IV
COMPARISON OF TRANSFER LEARNING PERFORMANCE WHEN PRETRAINING WITH DIFFERENT NUMBER OF SOURCE PATIENTS

# Source Patients	SBP MAE (mmHg)	DBP MAE (mmHg)
10	4.07	2.36
30	3.96	2.28
50	**3.84**	**2.24**
70	3.85	2.24
90	3.85	2.23

and 3.24 to 2.82 mmHg for SBP and DBP, respectively. While both approaches satisfy the AAMI standard, our transfer learning approach achieves the requirement by a larger margin.

Table III compares the transfer learning performance when different sets of network layers are finetuned using target patient data. We use the first 10 patients in our dataset as target patients for this experiment. The source model is pre-trained with the last 50 patients' data. These results are averaged over the 10 target patients. Evidently, retraining only the final convolution layer (Conv3) and fully connected layer (FC2) results in the best transfer learning performance. If the Conv3 layer is not personalized, the SBP MAE increases from 3.84 to 4.41 mmHg and the DBP MAE increases from 2.24 to 2.63 mmHg. This demonstrates the importance of personalizing the last convolutional layer in order to learn higher level features specific to the individual. One interesting observation is that, on average, it is better not to retrain the GRU with the target data. The average SBP and DBP MAEs when finetuning the GRU layer with the Conv3 and FC2 layer are 3.90 and 2.28 mmHg, respectively. If the GRU is not personalized, the average SBP and DBP MAEs are 3.84 and 2.24 mmHg, respectively. This may be because the GRU is modeling the temporal relationship between features, and not the features themselves. This indicates that the temporal modeling of PPG features is transferable across individuals in addition to the lower-level convolutional filters.

Table IV compares the transfer learning performance when different numbers of source patients are used for pretraining the initial model. Like the previous experiment, we use the first 10 patients in our dataset as target patients for this experiment and the results are averaged over these patients. We compare the transfer performance when using 10, 30, 50, 70, and 90 source patients for pretraining. We finetune the "Conv3, FC2" layer set during the transfer learning step. We observe that the MAEs for SBP and DBP decrease as more source patients are included but level off at 50 patients. The MAEs for SBP estimation when using 50, 70, and 90 source patients are 3.84, 3.85, and 3.85 mmHg, respectively. The MAEs for DBP estimation when using 50, 70, and 90 source patients are 2.24, 2.24, and 2.23 mmHg, respectively. These results demonstrate that including more than 50 source patients does not enhance the transfer learning performance. This indicates that there is sufficient variability and information among 50 patients to learn effective transferable features for PPG-BP estimation.

C. Effect of Training Set Size

Next, we discuss how our non-transfer and transfer learning performances change based on the number of target patient training samples. We test the model performance using 5 different amounts of personal training data: 3600, 1800, 360, 100, and 50 data samples. Since each input PPG segment is 5 seconds, 3600 samples correspond to 5 hours of data. For each case, the validation and test sets are kept the same in order to ensure a fair comparison. Figure 8 displays the relationship between the number of training samples and SBP (left) and DBP (right) estimation performance. The blue curves correspond to our non-transfer approach, namely BP-CRNN, while the red curves correspond to our transfer method, namely BP-CRNN-Transfer. Each point is labeled with the number of training samples and corresponding MAE. The black lines represent the performance of the dummy Mean Regressor, which always predicts the mean SBP and DBP of the target patient's training set. Again, we use the Mean Regressor's performance as a reference to ensure our model is learning more than simply estimating with the patient's mean SBP and DBP.

Evidently, using transfer learning improves performance for each number of training samples. As the number of training samples is reduced, the MAE increases for both approaches, but at a lower rate when utilizing transfer learning. When training with 100 data samples using the non-transfer approach, the MAE increases to 8.15 mmHg for SBP and 4.48 mmHg for DBP. In this case, the error is approaching that of the Mean Regressor, meaning the model has difficulty learning the PPG-BP relationship. If further reduced to 50 training samples, the model is unable to converge. This is why there is no point plotted for 50 samples when using our non-transfer approach. On the other hand, when using 100 training samples, the performance of our transfer learning approach for SBP and DBP is 5.52 and 3.38 mmHg, respectively. This corresponds to a 32.3% and 24.6% performance improvement for SBP and DBP estimation when

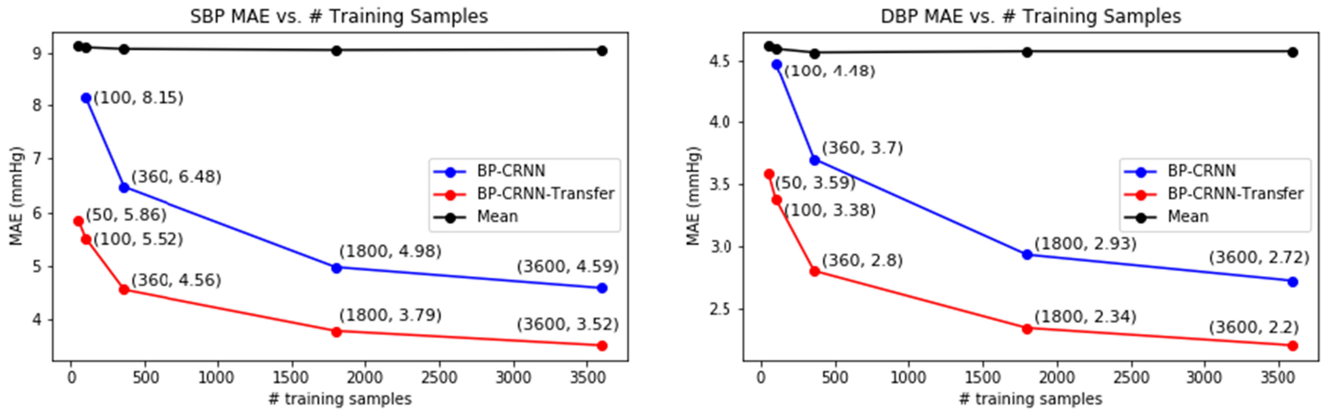


Fig. 8. BP estimation performance for different training set sizes. The labeled points for 360 and 3600 training samples indicate that our BP-CRNN-Transfer method can achieve equivalent performance to the non-transfer BP-CRNN method with 10× less data.

using our transfer learning technique. By comparing the non-transfer approach using 3600 samples to the transfer approach using 360 samples, we can see that the MAE is similar for SBP (4.59 vs. 4.56 mmHg) and DBP (2.72 vs. 2.80 mmHg) estimation. This indicates that 10× less personal PPG-BP data is required by our proposed transfer learning approach to achieve performance equivalent to that of a new personalized model trained with abundant data. For 50 training samples the model is able to converge using transfer learning, resulting in a MAE of 5.86 mmHg for SBP and 3.59 mmHg for DBP. The cuff-based standard is a MAE of $\leq$5 mmHg for both SBP and DBP [37]. Hence, our transfer learning technique satisfies this requirement for DBP and misses this requirement by 0.86 mmHg for SBP, when using 50 training samples. These results demonstrate that accurate personalized models can be trained even with limited personal PPG and BP data.

D. Bland-Altman and Correlation Analysis

Bland-Altman analysis is a technique for comparing a new measurement device or procedure to an approved method [38]. The goal is to assess the extent to which two methods designed to measure the same parameter are in agreement. Here, the two methods for BP measurement being compared include the invasive arterial catheter and our BP-CRNN-Transfer model. The difference in measurements between these two methods is plotted against the average measurement of the two devices. The difference between methods and mean of methods are calculated for each data sample using Eq. (7) and (8), respectively.

$$BP_{diff} = BP_{catheter} - BP_{BP-CRNN} \tag{7}$$

$$BP_{mean} = \frac{BP_{catheter} + BP_{BP-CRNN}}{2} \tag{8}$$

It is common to compute the 95% limits of agreement between measurement methods. These limits are defined as the average difference between measurement methods (blue dashed line in Figure 9) ± 1.96 $*$ standard deviation of the differences between measurement methods (red-dashed lines in Figure 9). For two methods to be considered comparable, Bland-Altman recommends that 95% of the samples should fall within these limits (red dashed lines). Among all 100 patients, 86% and 93% achieve this agreement for SBP and DBP measurement, respectively.

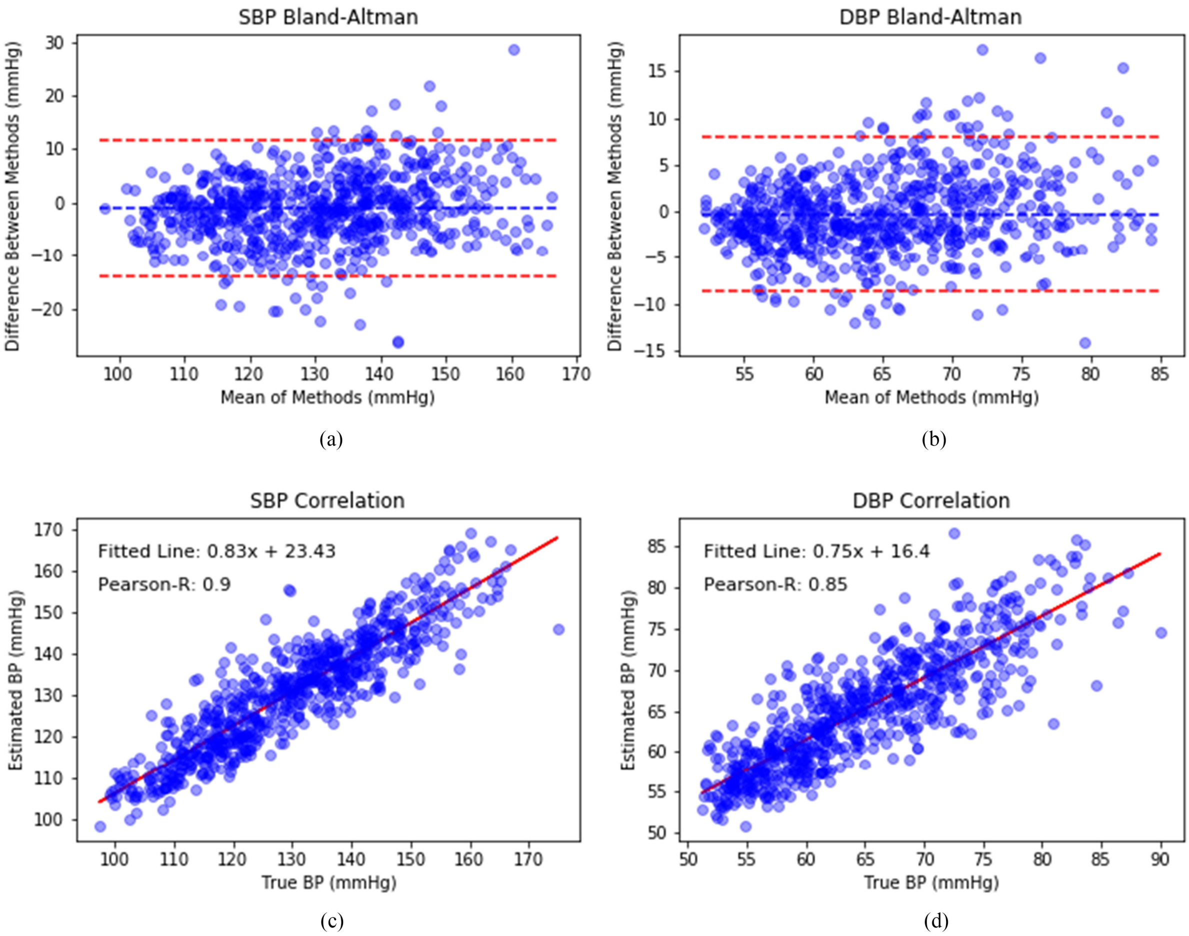
We also carry out Pearson correlation analysis [39] separately for each of the 100 patients to compare our method's estimated BP to the reference BP. The Pearson-R correlation coefficient is a measure of how linearly correlated two sets of data are. When using our non-transfer approach, the average and standard deviation of the Pearson-R coefficient is 0.83 ± 0.10 and 0.73 ± 0.17 for SBP and DBP, respectively. When using our transfer learning approach, the average and standard deviation of the Pearson-R coefficient is 0.90 ± 0.06 and 0.82 ± 0.12 for SBP and DBP, respectively. This increase in correlation again shows the ability of transfer learning to improve estimation performance.

Since it is not possible to show individual plots for each patient, we provide plots for one patient whose Pearson correlation is similar to the average correlation across all patients. Figure 9 displays both the Bland-Altman and correlation plots for SBP and DBP for this patient. 95.1% of the SBP differences and 95.6% of the DBP differences fall within the Bland-Altman limits of agreement. The correlation between estimated and reference BPs is 0.9 and 0.85 for SBP and DBP, respectively. These results demonstrate a high level of agreement between our model's estimated BP and the invasively measured BP from the arterial catheter.

E. Investigating Source Patient Selection

In this section, we discuss findings regarding source patient selection for individual target patients. Table IV compares results when using different numbers of source patients, however, these results represent an average and do not capture performance variations at the individual patient level. The goal of this experiment is to determine whether there are optimal smaller sets of source patients for individual target patients.

In order to determine the effect of using different source patients for individual target patients, multiple models are pre-trained. Table V displays the results for 3 different target patients, using 3 different pre-trained models for transfer learning. Model



(a)

(b)

(c)

(d)

Fig. 9. Bland-Altman and Pearson correlation analysis for one patient used to assess agreement between BP measurement methods. Plots (a) and (b) display Bland-Altman analysis for SBP and DBP, respectively. The red-dashed lines correspond to the average difference ± 1.96 $*$ standard deviation of differences. Plots (c) and (d) display the correlation between estimated and reference SBPs and DBPs, respectively.

TABLE V
TRANSFER PERFORMANCE OF DIFFERENT PRETRAINED MODELS.

	SBP MAE / DBP MAE (mmHg)		
	Patient 1	Patient 2	Patient 3
Model 1	4.73 / 3.27	7.96 / 4.99	**4.79 / 2.47**
Model 2	**4.47 / 3.18**	7.06 / 4.58	5.46 / 3.29
Model 3	4.76 / 3.20	**6.85 / 4.41**	5.05 / 3.12

1 represents the same initial model used in the previous experiments pre-trained with 50 source patients. Models 2 and 3 were pre-trained using different random sets of 10 source patients. For this experiment, 50 samples from the target patient are used to finetune each model.

On average, pretraining with 50 source patients (shown in Table IV) is better than pretraining with 10 source patients. However, for individual target patients, there may be certain smaller sets of source patients that result in better transfer learning performance, as shown in Table V. This performance increase can be significant, especially seen for Patient 2. Model 3 (pre-trained with 10 source patients) performs 13.9% and 11.6% better for SBP and DBP estimation compared to Model 1 (pre-trained with 50 source patients) for this target patient. These results indicate that transfer learning performance can be further improved by selecting a specific subset of source patients for individual target patients. In future work, we plan to investigate this idea of intelligent source patient selection for improving transfer learning performance.

IV. CONCLUSION

In this paper, we present an effective hybrid network architecture for personalized BP estimation using the PPG signal. In order to reduce the number of personal PPG-BP samples required for training, we provide a novel transfer learning approach that personalizes specific layers of the network. Our method is tested over a demographically diverse set of patients, and our estimation performance achieves the BHS and AAMI standards.

In this study, the training and inference are implemented on a personal computer. For future work, we will investigate a light-weight BP estimation model which can be implemented directly on a wearable device that collects PPG data while providing comparable performance to our current work. This will provide more real-time measurements and address concerns regarding data transmission and data privacy. BP measurement based on the PPG signal will enable a deeper understanding of how BP changes throughout the day, allowing the user to make adjustments in order to reach and maintain a healthy BP.

References

[1] C. Fryar, Y. Ostchega, C. Hales, G. Zhang, and D. Kruszon-Moran, "Hypertension prevalence and control among adults: United States, 2015–2016," *NCHS Data Brief*, vol. 289, pp. 1–8, Oct. 2017. [Online]. Available: https://www.cdc.gov/nchs/products/databriefs/db289.htm

[2] Centers for Disease Control and Prevention, National Center for Health Statistics, "Underlying cause of death 1999–2013 on CDC WONDER online database," CDC WONDER online database, 2015.

[3] The World Health Organization, "A global brief on hypertension: Silent killer, global public health crisis," World Health Day, Geneva, Switzerland, Apr. 2013.

[4] American College of Cardiology, "Guideline for the prevention, detection, evaluation, and management of high blood pressure in adults," *J. Amer. College Cardiol.*, vol. 71, no. 19, pp. 4–28, 2017.

[5] G. Ogedegbe and T. Pickering, "Principles and techniques of blood pressure measurement," *Cardiol. Clin.*, vol. 28, no. 4, pp. 571–586, 2010.

[6] J. Cho, "Current status and prospects of health-related sensing technology in wearable devices," *J. Healthcare Eng.*, vol. 2019, Jun. 2019, Art. no. 3924508.

[7] Y. Liang, Z. Chen, R. Ward, and M. Elgendi, "Photoplethysmography and deep learning: Enhancing hypertension risk stratification," *Biosensors*, vol. 8, no. 4, 2018, Art. no. 101.

[8] N. Hasanzadeh, M. M. Ahmadi, and H. Mohammadzade, "Blood pressure estimation using photoplethysmogram signal and its morphological features," *IEEE Sensors J.*, vol. 20, no. 8, pp. 4300–4310, Apr. 2020.

[9] M. Radha, K. De Groot, N. Rajani, C. C. P. Wong, N. Kobold, and V. Vos, "Estimating blood pressure trends and the nocturnal dip from photoplethysmography," *Physiol. Meas.*, vol. 40, no. 2, Jan. 2019, Art. no. 025006.

[10] G. Slapničar, M. Luštrek, and M. Marinko, "Continuous blood pressure estimation from PPG signal," *Informatica*, vol. 42, no. 1, pp. 33–42, 2018.

[11] S. G. Khalid, J. Zhang, F. Chen, and D. Zheng, "Blood pressure estimation using photoplethysmography only: Comparison between different machine learning approaches," *J. Healthcare Eng.*, vol. 2018, Oct. 2018, Art. no. 1548647.

[12] P. K. Lim, S.-C. Ng, N. H. Lovell, Y. P. Yu, M. P. Tan, and D. Mccombie, "Adaptive template matching of photoplethysmogram pulses to detect motion artefact," *Physiol. Meas.*, vol. 39, no. 10, pp. 1–12, Sep. 2018.

[13] J. Leitner, P. Chiang and S. Dey, "Personalized blood pressure estimation using photoplethysmography and wavelet decomposition," in *Proc. IEEE Int. Conf. E-health Netw., Application Serv. (HealthCom)*, Bogota, Colombia, 2019, pp. 1–6.

[14] P. Xia, J. Hu, and Y. Peng, "EMG-based estimation of limb movement using deep learning with recurrent convolutional neural networks: EMG-based estimation of limb movement," *Artif. Organs*, vol. 42, no. 5, pp. E67–E77, Oct. 2017.

[15] L. Fraiwan and K. Lweesy, "Neonatal sleep state identification using deep learning autoencoders," in *Proc. IEEE 13th Int. Colloq. Signal Process. Appl.*, Batu Ferringhi, 2017, pp. 228–231.

[16] U. R. Acharya, H. Fujita, O. S. Lih, Y. Hagiwara, J. H. Tan, and M. Adam, "Automated detection of arrhythmias using different intervals of tachycardia ECG segments with convolutional neural network," *Inf. Sci.*, vol. 405, pp. 81–90, Sep. 2017.

[17] G. Slapničar, N. Mlakar, and M. Luštrek, "Blood pressure estimation from photoplethysmogram using a spectro-temporal deep neural network," *Sensors*, vol. 19, no. 15, 2019, Art. no. 3420.

[18] O. Schlesinger, N. Vigderhouse, D. Eytan, and Y. Moshe, "Blood pressure estimation from PPG signals using convolutional neural networks and siamese network," in *Proc. ICASSP IEEE Int. Conf. Acoust., Speech Signal Process.*, Barcelona, Spain, 2020, pp. 1135–1139.

[19] C. Wang, F. Yang, X. Yuan, Y. Zhang, K. Chang, and Z. Li, "An end-to-end neural network model for blood pressure estimation using PPG signal," in *Artificial Intelligence in China*, Cham, Switzerland: Springer, pp. 262–272, 2020.

[20] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[21] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, South Brisbane, QLD, Australia, 2015, pp. 4580–4584.

[22] A. Johnson *et al.*, "MIMIC-III, a freely accessible critical care database," *Sci. Data*, 2016.

[23] A. Goldberger *et al.*, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, Jun. 2000.

[24] A. Johnson, T. Pollard, and R. Mark, "MIMIC-III clinical database (version 1.4)," *PhysioNet*, 2016. [Online]. Available: https://doi.org/10.13026/C2XW26

[25] A. Choi and H. Shin, "Photoplethysmography sampling frequency: Pilot assessment of how low can we go to analyze pulse rate variability with reliability?," *Physiol. Meas.*, vol. 38, no. 3, 2017, Art. no. 586.

[26] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. 13th Int. Conf. Artif. Intell. Statist.*, 2010, pp. 249–256.

[27] J. Chung, C. Gulcehre, K. Cho and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *Comput. Sci.*, Dec. 2014.

[28] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, pp. 448–456, 2015, *arXiv:1502.03167*.

[29] H. Chen, S. Lundberg, G. Erion, J. H. Kim, and S. Lee, "Deep transfer learning for physiological signals," 2020, *arXiv:2002.04770*.

[30] Y. Zhang and Q. Yang, "A survey on multi-task learning," in *Proc. IEEE Trans. Knowledge Data Eng.*, doi: 10.1109/TKDE.2021.3070203.

[31] A. Paszke *et al.*, "Automatic differentiation in PyTorch," 2017.

[32] D. Eigen, J. Rolfe, R. Fergus, and Y. LeCun, "Understanding deep architectures using a recursive convolutional network," *ICLR Workshop*, 2014.

[33] D. P. Kingma and J. L. Ba, "ADAM: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–41.

[34] L. Prechelt, "Early stopping—But when?" in *Neural Networks: Tricks of the Trade*, New York, NY, USA: Springer, pp. 53–67, 2012.

[35] E. O'Brien, B. Waeber, G. Parati, J. Stassen and M. Myers, "Blood pressure measuring devices: recommendations of the European society of hypertension," vol. 322, no. 7285, pp. 531–536, 2001.

[36] "Non-invasive sphygmomanometers—Part 2: Clinical investigation of automated measurement type," ANSI/AAMI/ISO 81060-2:2013, Arlington, VA, USA, 2016.

[37] Association for the Advancement of Medical Instrumentation (AAMI), "American national standard manual," *Electronic or Automated Sphygmomanometers, AASI/AAMI SP*, vol. 10, no. 2002, 2003.

[38] D. G. Altman and J. M. Bland, "Measurement in medicine: The analysis of method comparison studies," *Statistician*, vol. 32, no. 3, pp. 307–317, 1983.

[39] J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," in *Noise Reduction in Speech Processing*, New York, NY, USA: Springer, pp. 1–4, 2009.