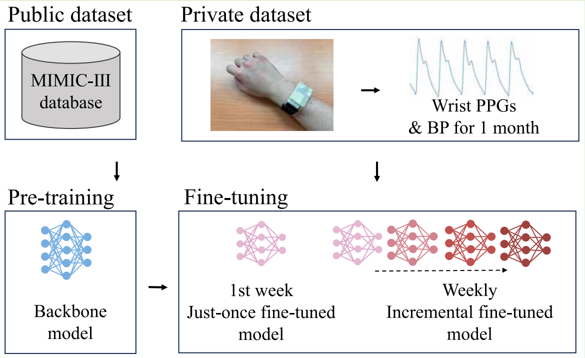
One-Month Evaluation of Blood Pressure Estimation Using a Fine-Tuned Model With Wristband-Type Photoplethysmograms

Satoshi Kamei, Suguru Kanoga, Masataka Yamamoto, *Member, IEEE*,
Hiroshi Takemura, *Member, IEEE*, and Mitsunori Tada


Public dataset
Private dataset
MIMIC-III database
Wrist PPGs & BP for 1 month
Pre-training
Fine-tuning
Backbone model
1st week Just-once fine-tuned model
Weekly Incremental fine-tuned model

***Abstract*—In the quest for a blood pressure (BP) measurement method that is applicable in daily life, deep learning-based BP estimation models utilizing photoplethysmograms (PPGs) have recently been introduced. To allow the routine deployment of BP estimation models that rely on wearable sensing PPGs, they should be able to sustain performance over an extended period. However, achieving stability is difficult due to variations in the distribution of PPG data caused by shifts in sensor positions and human conditions. This study assessed the performance of deep learning-based BP estimation models that were pretrained on a large-scale dataset and explored their adaptability and adjustment frequency during fine-tuning with wearable sensed small-scale data over a one-month period. Two distinct datasets were utilized: the publicly available Multiparameter Intelligent Monitoring in Intensive Care III Waveform Database Matched Subset version 1.0 and a private dataset recorded from 11 participants over four days (one day per week) using a customized wearable wristband-type PPG device. The results indicated the superior performance of the fine-tuned models compared with the non-fine-tuned models. Tuning the models every week resulted in greater performance than tuning only once in the first week. Balancing convenience and performance continues to be a practical challenge in this research field. Moreover, even the most accurate fine-tuned model does not consistently satisfy medical standards, presenting a hurdle for medical applications. This study revealed the effectiveness and adaptability of pretrained fine-tuned BP estimation models for long-term wearable sensing data; however, enhancing the estimation performance remains a challenge.**

***Index Terms*—Blood pressure (BP), deep learning, fine-tuning, long-term use, photoplethysmogram (PPG), wearable sensing.**

I. Introduction

IDENTIFYING and addressing signs of hypertension onset and assisting in preventing its development are crucial for maintaining well-being because hypertension is a significant cause of cardiovascular diseases, such as angina pectoris and myocardial infarction [1] and is a major cause of death due to stroke [2]. The number of patients with hypertension worldwide has doubled from 1990 to 2019 [3]. Hypertension can be identified when systolic blood pressure (SBP) is above 130 mmHg or when diastolic blood pressure (DBP) is above 80 mmHg [4]. SBP represents the blood pressure (BP) when the heart contracts and pumps blood from the heart to the body through the aorta. DBP represents the BP when the heart relaxes and receives blood from the body. The method most commonly used to measure BP in healthy people involves the use of a cuff-type sphygmomanometer on the upper arm, which is cumbersome and hampers routine BP measurements in daily life. The development of a simple and implicit BP measurement method may contribute to obtaining BP values with high temporal resolution, which would help to identify and address signs of early hypertension onset and could assist in preventing its development, resulting in the maintenance of well-being.

Manuscript received 19 April 2024; accepted 14 May 2024. Date of publication 29 May 2024; date of current version 1 July 2024. This work was supported in part by Japan Society for the Promotion of Science (JSPS) KAKENHI Grant-in-Aid for Scientific Research (B) under Grant JP23H03445 and Grant JP23K28135 and in part by the Japan Science and Technology Agency (JST)-Mirai Program under Grant JPMJMI21B1. The associate editor coordinating the review of this article and approving it for publication was Prof. Kea-Tiong (Samuel) Tang. *(Corresponding author: Suguru Kanoga.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Institutional Review Board of the National Institute of Advanced Industrial Science and Technology (AIST), Japan, under Application No. HF2022-1204, and performed in line with the Declaration of Helsinki.

Satoshi Kamei is with the Graduate School of Science and Technology, Tokyo University of Science (TUS), Chiba 278-8510, Japan, and also with the Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology (AIST), Tokyo 135-0064, Japan.

Suguru Kanoga and Mitsunori Tada are with the Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology (AIST), Tokyo 135-0064, Japan (e-mail: s.kanouga@aist.go.jp).

Masataka Yamamoto is with the Graduate School of Science and Technology, Tokyo University of Science (TUS), Chiba 278-8510, Japan, and also with the Graduate School of Advanced Science and Engineering, Hiroshima University, Hiroshima 739-8527, Japan.

Hiroshi Takemura is with the Graduate School of Science and Technology, Tokyo University of Science (TUS), Chiba 278-8510, Japan.

Digital Object Identifier 10.1109/JSEN.2024.3404355

To develop a routinely applicable BP measurement method, machine-learning and deep-learning models using data obtained from photoplethysmograms (PPGs) to estimate BP values have been proposed [5], [6], [7], [8], [9], [10], [11], [12]. Such studies have focused on constructing advanced models using large-scale datasets collected from traditional nonwearable PPG devices used in hospitals. Furthermore, they reported an increasing number of models that exhibited minimal errors between the true and estimated BP values, meeting the standards for estimating BPs. PPGs provide an optical measurement of subcutaneous blood flow, making them well suited for implementation in wearable devices because of the affordability and lightweight nature of the sensors. Therefore, such devices have the potential to be simple and implicit BP measurement tools.

Recently, personalization and continual learning have become common in deep-learning-based BP estimation for PPGs [13], [14], [15], [16], [17]. Both are important keywords, but they have used separately. The studies developing personalization methods based on fine-tuning [13], [14], [15], [16] partition large-scale or small-scale datasets collected from nonwearable or wearable devices into source and target subsets. However, fine-tuning is conducted only once, and there is a lack of continual learning. Alternatively, a study developing continual learning methods [17] addresses the challenge of training the model with a small initial dataset and gradually adding data to update the model. It does not pretrain a model with a large-scale dataset. In studies that use PPGs to develop routinely applicable BP estimations, customized wearable PPG equipment can initially yield only small datasets; however, several large-scale datasets are publicly available [18], [19]. If a backbone (BB) model is pretrained using such a dataset and incrementally updated using a small-scale dataset collected from wearable devices, it may be useful over time and can effectively track hypertension development.

Thus, in this study, we investigated the performance of deep-learning BP estimation models that were pretrained using a large-scale dataset and were then fine-tuned with data from a wearable device over a one-month period. The routine use of BP estimation models based on wearable PPG devices requires stable estimation performance over the long term; however, the distribution of PPG data can vary with changes in sensor positions [20] and in human conditions [21]. In addition, it is unclear how well the pretrained and fine-tuned BP estimation models perform when they are adjusted and how often they should be adjusted. We compared two types of fine-tuned models.

1) A model in which the pretrained model was fine-tuned only once with data from the first day of the long-term period, called the just-once fine-tuned (JO-FT) model.
2) An incremental model in which the pretrained model was incrementally fine-tuned with data from every measurement day, called the incremental fine-tuned (INC-FT) model.

To evaluate the performance of the pretrained and fine-tuned models, we used two datasets: a large-scale public dataset, called the Multiparameter Intelligent Monitoring in Intensive Care III Waveform Database Matched Subset version 1.0 (MIMIC-III dataset) [19], and a small-scale private dataset recorded from 11 participants on four days within a one-month period using our customized wearable wristband-type device.

A deep-learning BP estimation model using a large-scale dataset from a hospital intensive care unit (ICU) was fine-tuned to a wearable sensed small-scale dataset from healthy participants to confirm the improvement in accuracy over time. The two types of fine-tuned models were more accurate than models trained on the respective datasets alone and met the Advancement of Medical Instrumentation (AAMI) [22] and British Hypertension Society (BHS) [23] standards except for some long-term validation measurement dates. In particular, the INC-FT model achieved the best performance. The results suggest that pretraining with a large-scale dataset and fine-tuning with a small-scale dataset are effective in deep learning-based BP estimation for PPGs and that performance can be improved by incorporating data that increase incrementally through wearable sensing. This study represents a significant contribution to the development of an implicit BP measurement method that can be readily applied in routine well-being monitoring.

II. Materials and Methods

In this study, three deep learning-based BP estimation architectures (i.e., PP-Net [7], modified PP-Net (mPP-Net) [9], and spectral–temporal residual network (ST-ResNet) [5]), described in our previous study [24] were used as the BB models. These models were then fine-tuned with a small amount of calibration data from a private dataset measured using a customized wearable wristband-type PPG device to assess the feasibility of using the fine-tuned model over a one-month period. The details of the datasets, pretraining/fine-tuning methods, and evaluation procedures are described below.

A. Public Dataset (MIMIC-III Dataset)

To train the BB deep learning-based BP estimation models, the MIMIC-III dataset published by PhysioNet [19] was used. This dataset contains physiological parameters and biological signals measured in approximately 30 000 patients in the ICU. PPG data and BP were measured using bedside monitors at a sampling rate of 125 Hz. Signals were measured using a pulse oximeter attached to the fingertip and a catheter inserted directly into the blood vessels.

The processing stream used to create the training dataset is shown in Fig. 1. Waveform and numerical data were collected from the MIMIC-III Waveform Database Matched Subset. The physiological signals, such as the PPG and arterial BP, were recorded from the waveform data. Then, files with too little data, files with physiological signals missing for more than 10 min in the waveform data, and files with missing BP or heart rate in the numerical data were removed. The collected data were converted to the MATLAB native format (i.e., *.mat file) with the WaveForm DataBase Toolbox.

After data collection and cleaning, the PPG data and BPs were segmented into 8-s intervals with a 6-s overlap using a sliding window approach, according to the previous studies of PPG-based BP estimation [7], [9], [25]. Abnormal segments were removed by setting range thresholds for the SBP and DBP values (i.e., SBP $\geq$ 260, SBP $\leq$ 60, DBP $\geq$ 125, and

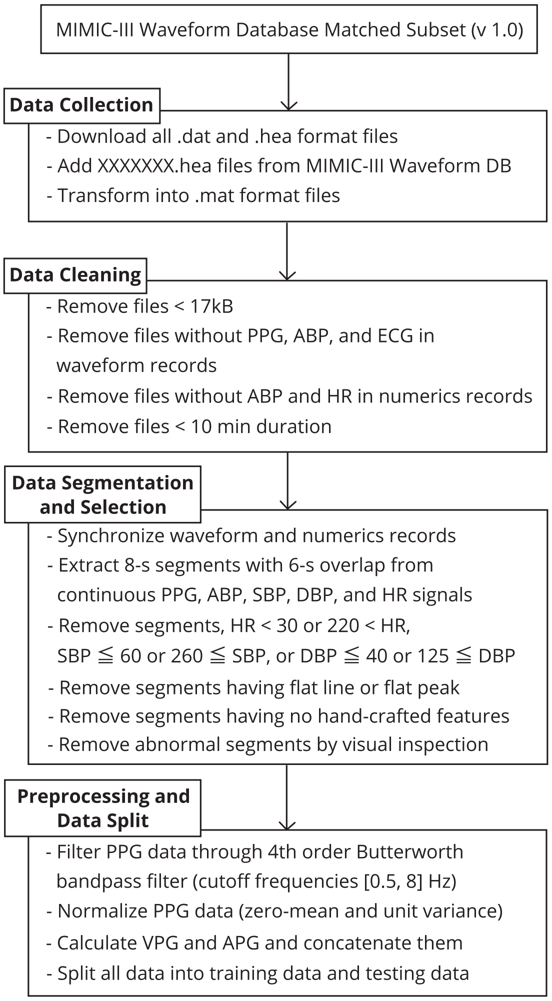

MIMIC-III Waveform Database Matched Subset (v 1.0)
Data Collection
- Download all .dat and .hea format files
- Add XXXXXXX.hea files from MIMIC-III Waveform DB
- Transform into .mat format files
Data Cleaning
- Remove files < 17kB
- Remove files without PPG, ABP, and ECG in waveform records
- Remove files without ABP and HR in numerics records
- Remove files < 10 min duration
Data Segmentation and Selection
- Synchronize waveform and numerics records
- Extract 8-s segments with 6-s overlap from continuous PPG, ABP, SBP, DBP, and HR signals
- Remove segments, HR < 30 or 220 < HR, SBP ≦ 60 or 260 ≦ SBP, or DBP ≦ 40 or 125 ≦ DBP
- Remove segments having flat line or flat peak
- Remove segments having no hand-crafted features
- Remove abnormal segments by visual inspection
Preprocessing and Data Split
- Filter PPG data through 4th order Butterworth bandpass filter (cutoff frequencies [0.5, 8] Hz)
- Normalize PPG data (zero-mean and unit variance)
- Calculate VPG and APG and concatenate them
- Split all data into training data and testing data

Fig. 1. Processing stream up to the creation of the dataset modified from [24].

DBP $\leq$ 40) [7], [26]. The remaining segments were filtered by a fourth-order Butterworth filter with cutoff frequencies of 0.5 and 8 Hz [5]. In addition, the PPG data were normalized to a zero mean and unit variance. Based on the segments, the first and second derivatives [i.e., velocity PPG (VPG) and acceleration PPG (APG)] were calculated and concatenated to the PPG segments, resulting in 3-D 8-s segments [i.e., the shape of a segment is (3, 1000)]. The data of 1811 participants were split into 80% training patients (1453 patients) and 20% testing patients (358 patients). In the subject-dependent case, 80% of the data from each of the 358 testing patients were added to the training set (see Fig. 2). In this study, we did not use the testing data $\boldsymbol{X}_{\text{test}}$ in the figure because only the training data were needed for pretraining BB models. To reduce the computational cost, the segmented signals were down-sampled using a scaling factor of 4 [i.e., the shape of a segment was (3, 250)] [7].

B. Pretraining Backbone (BB) Model

PP-Net [7], mPP-Net [9], and ST-ResNet [5] were pretrained as BB deep learning-based BP estimation models using 3-D 8-s segments of the MIMIC-III dataset. This model is referred to as the BB model. PP-Net and mPP-Net process PPG, VPG, and APG signals using two and three convolutional layers and two long short-term memory layers (see [24, Fig. 5]). ST-ResNet processes PPG, VPG, and APG signals using an ST block and five residual blocks. The ST block captured spectral and temporal representations by transforming a signal into a spectrogram using a gate recurrent unit (GRU) layer. The residual block had two streams: 1) a shortcut 1-D convolutional layer and 2) three 1-D convolutional layers. After the residual blocks, the learned features were further processed using a GRU layer. The feature vectors from the ST block, residual blocks, and GRU layer were concatenated for two dense layers [24, Fig. 6]. In this study, the models were trained without normalizing the output values.

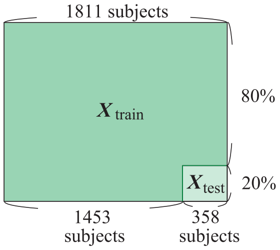


Fig. 2. Training and testing data in the subject-dependent case modified from [24].

TABLE I
LOSS FUNCTION, BATCH SIZE, INITIAL LEARNING RATE, STEP SIZE, AND DROP LEVEL OF LEARNING RATE DECAY, NUMBER OF EPOCHS, OPTIMIZER, AND NUMBER OF TRAINABLE PARAMETERS IN THREE DEEP-LEARNING-BASED BP ESTIMATION ARCHITECTURES

	PP-Net	mPP-Net	ST-ResNet
Loss function	MSE	Huber	L1
Batch size	100	100	100
Initial learning rate	0.001	0.001	0.001
Step size	50	50	10
Drop level	0.1	0.1	0.1
# of epochs	200	200	30
Optimizer	Adam	Adam	RMSprop
# of trainable parameters	187,431	398,883	1,064,207

The BB models were implemented on the PyTorch platform version 1.10.0. The loss function, batch size, initial learning rate, step size, and drop level of the learning rate decay, number of epochs, optimizer, and number of trainable parameters of the three deep-learning-based BP estimation architectures are shown in Table I. These conditions are the same as those used in a previous study [24].

C. Private Dataset

To assess the performance of the fine-tuned deep learning-based BP estimation model in a real-life environment over a long period, we measured a private dataset from healthy participants using a customized wearable wristband-type PPG device. SBP, DBP, and PPG were measured in 11 healthy participants (eight men, age: 31.1 ± 4.3 years and three women, age: 31.0 ± 11.3 years). This dataset has been made publicly available at: https://github.com/aistairc/One_Month_Wrist_PPG_Dataset. Each participant participated in the experiment approximately weekly for a total of four days (within approximately one month). Each day, the participants lay on their backs in a relaxed position in a row of well-cushioned chairs [see Fig. 3(a)], during which the PPG was measured for six min

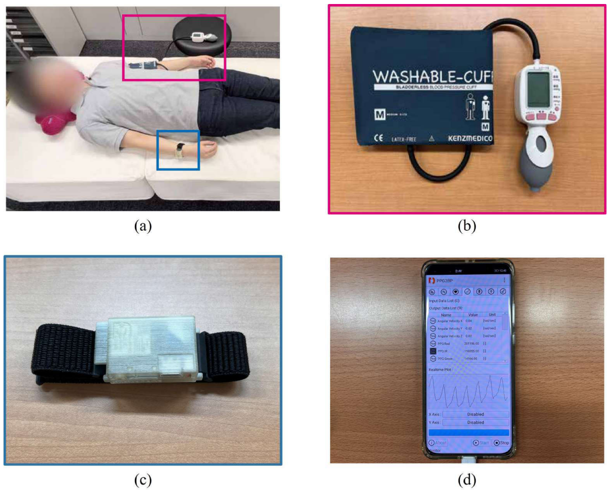


Fig. 3. (a) Experimental status of a participant during PPG and BP measurements. (b) Cuff-type sphygmomanometer. (c) Customized wearable wristband-type PPG device. (d) Self-made smartphone application that communicates with the PPG device and collects data.

in two sessions. In a previous study, the deep learning model was able to converge using transfer learning when 50 samples of 5-s windows without overlap (i.e., 250-s calibration data) were prepared [14]. Based on this study, we decided to set the length of one session to 6 min, because we could provide 240-s calibration data for fine tuning. SBP and DBP were measured in the left upper arm after each 6-min PPG session using a cuff-type sphygmomanometer (KM-370 II, Kenzmedico Company Ltd., Honjo, Japan), as shown in Fig. 3(b). As cuff-type sphygmomanometers apply pressure to the measurement area during BP measurement, which changes the condition of the blood vessels, a 2-min break was taken after each measurement, and a second 6-min PPG measurement was taken after the vascular condition had returned to normal. The customized wristband-type PPG device shown in Fig. 3(c) consists of a microcontroller (ESP32-WROOM-32D, Espressif Systems Pte. Ltd., Shanghai, China) and a pulse oximeter (MAX30101, Maxim Integrated, San Jose, CA, USA), which were housed in a 3-D printed case. PPG measurements were conducted at a sampling rate of 200 Hz from the right wrist using a pulse oximeter. The data were continuously sent to a smartphone (ELS-NX9, Huawei Technologies Company Ltd., Shenzhen, China), as shown in Fig. 3(d), via Wi-Fi by the microcontroller. This study was approved by the Institutional Review Board of the National Institute of Advanced Industrial Science and Technology (AIST), Japan (HF2022-1204). All participants provided their informed consent for the content and risks of the experiment as well as the open sourcing of the data prior to starting the experiment.

In addition to the MIMIC-III dataset, 6-min PPG measurements were segmented into 8-s intervals with a 6-s overlap using the sliding window approach. No abnormal segments for SBP or DBP values in this experiment were noted because all the participants were healthy. All the segments were filtered using a fourth-order Butterworth filter with cutoff frequencies of 0.5 and 8 Hz. An example of a filtered 8-s PPG signal is shown in Fig. 4. In addition, the PPG values were normalized to a zero mean and unit variance. Based on the PPG segments, the VPGs and APGs were calculated and concatenated to the PPG segments, resulting in 3-dimensional 8-s segments. Here, the shape of the concatenated segment was (3, 1600) because the sampling rate was 200 Hz. To match the shape to the MIMIC-III dataset, the segmented signals were downsampled by a scaling factor of 8 after cubic spline interpolation [i.e., the shape of a segment was (3, 250)]. No abnormal segments in SBP or DBP values were present; however, some noisy data were present due to the use of wearable sensors. To determine whether the segment was clean or noisy, we extracted 46-D handcrafted features containing the amplitude, time, and area features described in a previous study [24]. If even one of the 46 features could not be extracted, the segment was determined to be noisy and eliminated from the dataset. Through this segment quality-check process, the average number of segments in each session and each participant decreased from 175.0 ± 2.0 to 77.0 ± 64.8. The remaining segments were assigned duplicate SBP and DBP values that were measured every 6 min.

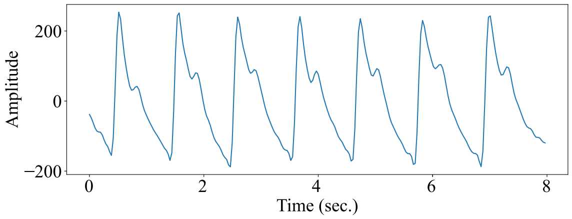


Fig. 4. Example of a filtered 8-s PPG signal.

D. Full-Scratch (FS) Model

To compare the performance of the fine-tuned pretrained models, the same architectures described in Section II-B were fully scratched using a small amount of calibration data obtained during the first section of each day in the private dataset measurement using the customized wearable wristband-type PPG device (see Fig. 5). The hyperparameters for this model, except for the initial learning rate, were the same as those described in Table I. The initial learning rate was set to 0.01 due to the small amount of data.

E. Fine-Tuning BB Model (JO-FT and INC-FT Models)

The BB models described in Section II-B were fine-tuned using a small amount of calibration data. Fine-tuning can transfer the represented knowledge of a model pretrained on a large dataset to a new dataset by updating the parameters of some or all of the layers in the pretrained model, even if the new dataset is small [27]. Therefore, we attempted to transfer the represented knowledge of the PP-Net, mPP-Net, and ST-ResNet models trained with the MIMIC-III dataset to our private dataset by updating the parameters of all the layers in the BB models with calibration data. Two fine-tuning approaches were used: 1) fine-tuning using calibration data from only the first week and 2) incremental fine-tuning using the calibration data from each week (see Fig. 5). These models are hereafter referred to as the JO-FT model and the INC-FT model. The hyperparameters for these models, except for the learning rate decay step size and number of epochs, were the same as those described in Table I. The learning rate and number of epochs were set to 10 and 30, respectively.

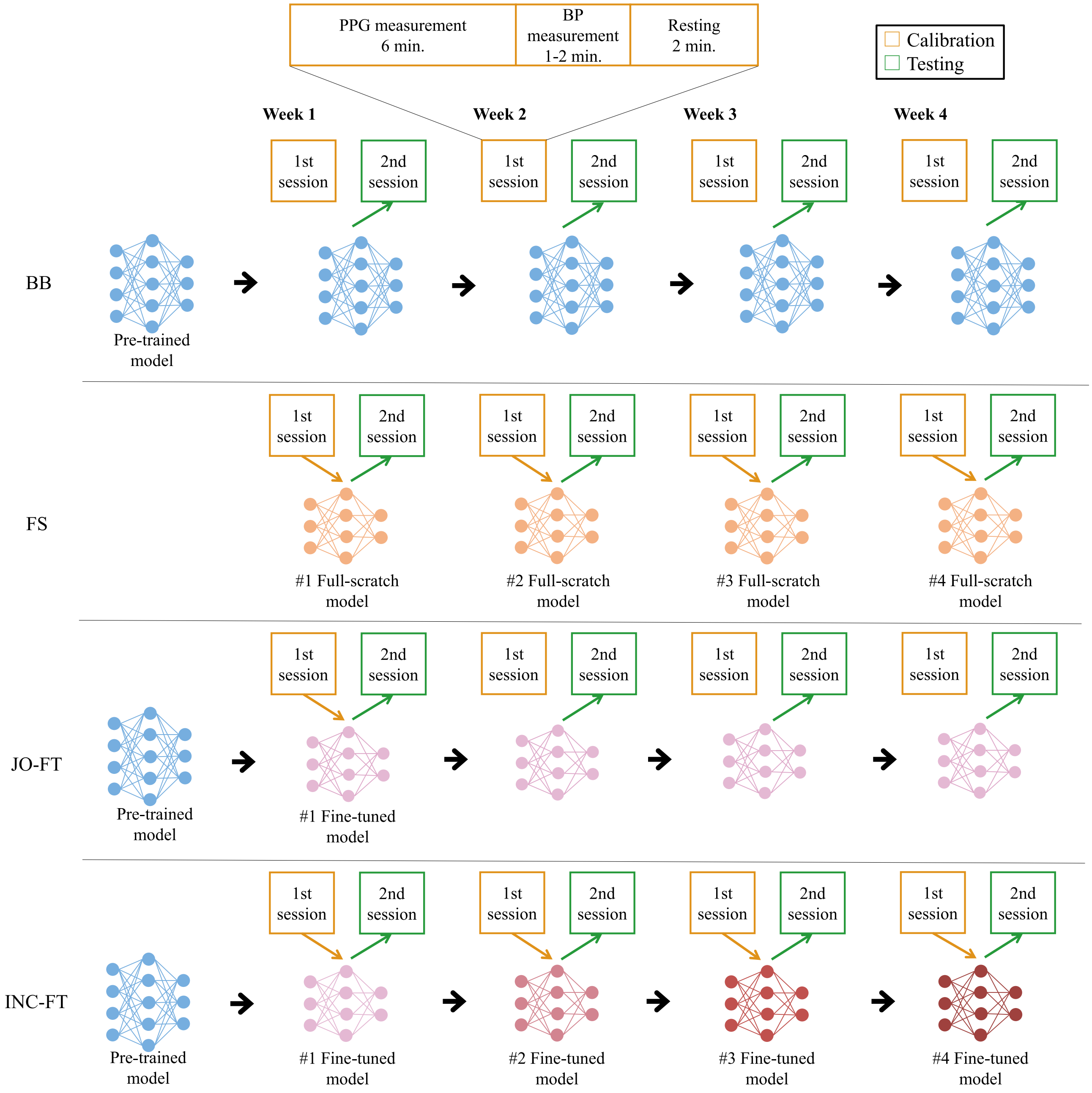


Fig. 5. Four training, tuning, and testing streams for the BB, FS, JO-FT, and INC-FT models.

F. Evaluation

To evaluate the performance of the fine-tuned deep learning-based BP estimation models over a one-month period with wrist PPG data, we compared the performances of the BB, FS, JO-FT, and INC-FT models of three deep learning architectures. In all models, data from the second session on all days were used as the testing data and the mean absolute errors (MAEs) between the true and estimated BP values were calculated. Multiple group comparisons across the three models based on MAEs were examined using the Kruskal–Wallis test, and paired comparisons among the three models were performed using the Dwass–Steel–Critchlow–Fligner test. In addition, the mean, variance, and range of the estimated values were calculated to verify the diversity of the output values of the models. Violin plots of the predicted and true BP values in the calibration and testing sessions are presented as a qualitative evaluation. Furthermore, we examined whether the accuracy of BP estimation met the AAMI and BHS standards, which are highly regarded in the medical field. The AAMI standard assesses performance based on whether the mean error (ME) for SBP and DBP is below 5 mmHg and the standard deviation (SD) of the error is below 8 mmHg. The BHS standard defines the specific performance requirements for each grade, as listed in Table II. Grade A represents the highest level of accuracy that modeling studies strive to achieve. Abnormal values were estimated for some of the segments. Thus, a postprocessing thresholding process was used to exclude segments for which the estimates were abnormal in one of the models (i.e., SBP $\geq$ 260, SBP $\leq$ 60, DBP $\geq$ 125, and DBP $\leq$ 40). The distributions of SBPs and

TABLE II
BHS GRADING SCALE FOR BP ESTIMATION

Grade	Cumulative absolute difference between true and estimated (%)		
	≤ 5 (mmHg)	≤ 10 (mmHg)	≤ 15 (mmHg)
A	60	85	95
B	50	75	90
C	40	65	85

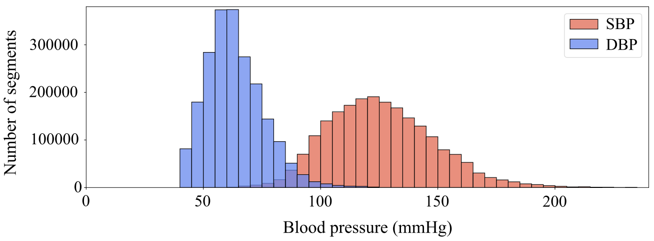


Fig. 6. Distributions of SBP and DBP in pretraining data for the 1811 participants of the public dataset after excluding segments for which estimates were abnormal in any one model (i.e., SBP ≥ 260, SBP ≤ 60, DBP ≥ 125, and DBP ≤ 40).

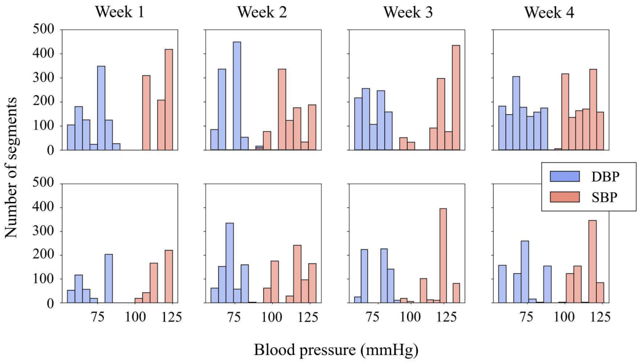


Fig. 7. Distributions of SBPs and DBPs in first (top) and second (bottom) sessions for the 11 participants in each week of the private dataset after excluding segments for which estimates were abnormal in any one model (i.e., SBP ≥ 260, SBP ≤ 60, DBP ≥ 125, and DBP ≤ 40).

DBPs in the pretraining data for the 1811 participants of the public dataset and in the first and second sessions for the 11 participants each week of the private dataset, excluding segments with abnormally estimated values, are shown in Figs. 6 and 7.

III. RESULTS

The results of the Kruskal–Wallis test across the four models based on the MAEs with three deep learning architectures are shown in Table III. In all architectures, both SBP and DBP were significantly different among the four models for all weeks ($p < 0.01$). The MAEs of the BB, FS, JO-FT, and INC-FT models over four weeks and the results of the Dwass–Steel–Critchlow–Fligner test for the four models are listed in Table IV. The results of the JO-FT and INC-FT models at week 1 were the same because the conditions were the same. FS models were more accurate than the BB models, confirming that within-dataset analysis can show better performance than cross-dataset analysis. The fine-tuned models (i.e., the JO-FT and INC-FT models) exhibited smaller MAEs than did the BB and FS models. Both SBP and DBP differed significantly between the fine-tuned and BB models and between the fine-tuned and FS models for all weeks ($p < 0.05$), except for week 4 of the JO-FT model based on PP-Net for SBP estimation. The INC-FT model had the smallest MAEs for SBP and DBP for all weeks among the four models except for week 3 of the JO-FT model based on mPP-Net for both SBP and DBP estimation and week 4 of the JO-FT model based on ST-ResNet for DBP estimation. Of the three architectures, ST-ResNet had the best average performance in the FS, JO-FT, and INC-FT models for both SBP and DBP estimation. Thus, in this study, ST-ResNet will be treated henceforth.

TABLE III
RESULTS OF THE KRUSKAL–WALLIS TEST ACROSS THE FOUR MODELS BASED ON MAEs WITH THREE DEEP LEARNING ARCHITECTURES

Architecture	Week	SBP		DBP	
		Kruskal–Wallis H	p value	Kruskal–Wallis H	p value
PP-Net	1	H(3, 1347) = 1042.0	< 0.01	H(3, 1347) = 170.7	< 0.01
	2	H(3, 2310) = 1142.4		H(3, 2310) = 218.7	
	3	H(3, 1884) = 858.6		H(3, 1884) = 233.7	
	4	H(3, 2151) = 1365.8		H(3, 2151) = 83.4	
	Avg.	H(3, 7701) = 3923.3		H(3, 7701) = 600.9	
mPP-Net	1	H(3, 1347) = 1211.4	< 0.01	H(3, 1347) = 256.9	< 0.01
	2	H(3, 2310) = 1512.1		H(3, 2310) = 276.9	
	3	H(3, 1884) = 1087.3		H(3, 1884) = 446.8	
	4	H(3, 2151) = 1333.3		H(3, 2151) = 89.8	
	Avg.	H(3, 7701) = 4726.4		H(3, 7701) = 764.8	
ST-ResNet	1	H(3, 1347) = 1112.4	< 0.01	H(3, 1347) = 1105.5	< 0.01
	2	H(3, 2310) = 1650.4		H(3, 2310) = 1716.6	
	3	H(3, 1884) = 1520.9		H(3, 1884) = 1399.0	
	4	H(3, 2151) = 1478.3		H(3, 2151) = 1263.3	
	Avg.	H(3, 7701) = 5603.6		H(3, 7701) = 5464.6	

The means, SDs, and ranges of the true and estimated SBP and DBP values from the BB, FS, JO-FT, and INC-FT models based on ST-ResNet over the four weeks are shown in Table V. In addition, violin plots of the true and estimated SBP and DBP values from the BB, FS, JO-FT, and INC-FT models based on ST-ResNet over the four weeks are shown in Fig. 8. In our dataset, the true values ranged from 95 to 133 mmHg for SBP and from 55 to 87 mmHg for DBP; however, the BB model estimated BP values that were far from the mark (71.45–184.91 mmHg for SBP and 40.05–124.37 mmHg for DBP). In contrast, the estimated BP values from the FS, JO-FT, and INC-FT models were close to the range of true values, with few segments showing extremely high or low values (FS model: 75.45–165.89 mmHg for SBP and 45.23–104.91 mmHg for DBP; JO-FT model: 91.57–165.89 mmHg for SBP and 51.58–120.78 mmHg for DBP; and INC-FT model: 86.02–127.83 mmHg for SBP and 52.77–98.34 mmHg for DBP).

A comparative analysis of the AAMI standards for the BB, FS, JO-FT, and INC-FT models based on ST-ResNet over the four-week period is presented in Table VI. The AAMI standard requires that the ME between the true and estimated SBP/DBP values be less than 5 mmHg and that the SD of the error be less than 8 mmHg in at least 85 participants [22]. Although only 11 participants were included in the study, the AAMI standard was used as a reference for evaluation. On average, across all weeks, the JO-FT and INC-FT models met the AAMI standard, whereas the other models did not. The ME and SD of the error estimated using the BB model deviated from the reference criterion of the AAMI standard for all weeks. In the FS model, SBP did not meet the criteria, and DBP met the criteria in two of the weeks. In the JO-FT model, SBP

TABLE IV
MAEs of BB, FS, JO-FT, and INC-FT Models Based on Three Deep-Learning Architectures Over Four Weeks and Results of the Dwass–Steel–Critchlow–Fligner Test ($p \geq 0.05$: ns, $0.01 \leq p < 0.05$: *, and $p < 0.01$: **). The Results Shown in Bold Are the Best Results of the Week in the Four Models

Architecture	Model	Week	SBP					DBP				
			MAE (mmHg)	vs BB	vs FS	vs JO-FT	vs INC-FT	MAE (mmHg)	vs BB	vs FS	vs JO-FT	vs INC-FT
PP-Net	BB	1	22.22	–	**	**	**	8.03	–	**	**	**
		2	27.15	–	**	**	**	7.47	–	ns	**	**
		3	18.20	–	**	**	**	9.47	–	ns	**	**
		4	25.26	–	**	**	**	8.39	–	**	ns	*
		Avg.	23.59	–	**	**	**	8.32	–	**	**	**
	FS	1	10.79	**	–	**	**	9.22	**	–	**	**
		2	16.38	**	–	**	**	7.75	ns	–	**	**
		3	14.46	**	–	**	**	9.55	ns	–	**	**
		4	8.87	**	–	ns	**	9.89	**	–	**	**
		Avg.	12.74	**	–	**	**	9.07	**	–	**	**
	JO-FT	1	**3.96**	**	**	–	–	**4.90**	**	**	–	–
		2	8.73	**	**	–	ns	5.14	**	**	–	ns
		3	7.26	**	**	–	ns	5.44	**	**	–	ns
		4	9.42	**	ns	–	**	7.52	ns	**	–	**
		Avg.	7.76	**	**	–	**	5.74	**	**	–	ns
	INC-FT	1	**3.96**	**	**	–	–	**4.90**	**	**	–	–
		2	**8.25**	**	**	ns	–	**4.66**	**	**	ns	–
		3	**6.41**	**	**	ns	–	**5.16**	**	**	ns	–
		4	**6.09**	**	**	**	–	**7.17**	*	**	**	–
		Avg.	**6.44**	**	**	**	–	**5.56**	**	**	ns	–
mPP-Net	BB	1	28.66	–	**	**	**	11.00	–	*	**	**
		2	32.78	–	**	**	**	7.65	–	ns	**	**
		3	22.14	–	**	**	**	8.00	–	**	**	**
		4	33.63	–	**	**	**	11.99	–	*	**	**
		Avg.	29.76	–	**	**	**	9.59	–	ns	**	**
	FS	1	9.15	**	–	**	**	8.83	*	–	**	**
		2	17.50	**	–	**	**	6.31	ns	–	**	**
		3	14.20	**	–	**	**	11.24	**	–	**	**
		4	8.30	**	–	*	**	10.28	*	–	*	**
		Avg.	12.56	**	–	**	**	9.10	ns	–	**	**
	JO-FT	1	**2.53**	**	**	–	–	**4.95**	**	**	–	–
		2	8.67	**	**	–	**	5.20	**	**	–	ns
		3	**5.66**	**	**	–	ns	**3.70**	**	**	–	**
		4	9.35	**	*	–	**	9.28	**	*	–	**
		Avg.	7.10	**	**	–	**	6.00	**	**	–	ns
	INC-FT	1	**2.53**	**	**	–	–	**4.95**	**	**	–	–
		2	**7.11**	**	**	**	–	**4.07**	**	**	ns	–
		3	5.81	**	**	ns	–	5.24	**	**	**	–
		4	**7.08**	**	**	**	–	**7.99**	**	**	**	–
		Avg.	**6.00**	**	**	**	–	**5.66**	**	**	ns	–
ST-ResNet	BB	1	49.60	–	**	**	**	38.18	–	**	**	**
		2	51.50	–	**	**	**	35.80	–	**	**	**
		3	47.90	–	**	**	**	35.57	–	**	**	**
		4	48.64	–	**	**	**	34.26	–	**	**	**
		Avg.	49.50	–	**	**	**	35.78	–	**	**	**
	FS	1	7.45	**	–	**	**	7.78	**	–	**	**
		2	11.80	**	–	**	**	5.96	**	–	**	**
		3	7.74	**	–	**	**	6.62	**	–	**	**
		4	6.73	**	–	**	**	8.91	**	–	**	ns
		Avg.	8.63	**	–	**	**	7.26	**	–	**	**
	JO-FT	1	**4.10**	**	**	–	–	**4.15**	**	**	–	–
		2	7.51	**	**	–	ns	5.88	**	**	–	**
		3	9.32	**	**	–	**	4.91	**	**	–	ns
		4	4.72	**	**	–	*	**7.15**	**	**	–	**
		Avg.	6.62	**	**	–	**	5.65	**	**	–	ns
	INC-FT	1	**4.10**	**	**	–	–	**4.15**	**	**	–	–
		2	**7.78**	**	**	ns	–	**2.96**	**	**	**	–
		3	**4.90**	**	**	**	–	**4.57**	**	**	ns	–
		4	**4.07**	**	**	*	–	8.38	**	ns	**	–
		Avg.	**5.39**	**	**	**	–	**5.07**	**	**	ns	–

met the criteria for two weeks, and DBP met the criteria for three weeks. In the INC-FT model, both SBP and DBP met the criteria for three of the four weeks.

The results of a comparative analysis of the BHS standards for BB, FS, JO-FT, and INC-FT models based on ST-ResNet over the four weeks are summarized in Table VII. The BHS standard requires Grade A used in medical settings [23]. Even the INC-FT model did not satisfy this requirement. The results based on the BHS standard were less than Grade C for both SBP and DBP in all weeks in the BB model. The average

TABLE V
MEAN, SD, AND RANGE OF TESTING/TRUE AND ESTIMATED BP VALUES FROM THE BB, FS, JO-FT, AND INC-FT MODELS BASED ON ST-RESNET OVER A FOUR-WEEK PERIOD

	Week	SBP (mmHg)			DBP (mmHg)		
		Mean	S.D.	Range	Mean	S.D.	Range
True	1	117.12	5.73	104.00 – 124.00	71.65	9.57	55.00 – 82.00
BB		148.08	38.55	75.56 – 183.93	95.31	28.96	40.05 – 118.69
FS		111.55	5.96	96.87 – 165.89	69.71	3.87	61.54 – 104.91
JO-FT INC-FT		114.04	4.68	102.73 – 123.94	72.33	7.67	54.72 – 98.34
True	2	114.46	10.39	95.00 – 127.00	70.95	6.38	59.00 – 86.00
BB		142.21	41.28	75.22 – 184.91	90.59	31.27	40.17 – 119.13
FS		104.44	6.64	92.01 – 122.29	67.68	3.62	59.81 – 77.56
JO-FT		112.92	7.45	96.28 – 164.22	68.80	7.87	51.58 – 120.78
INC-FT		106.92	8.01	86.02 – 120.73	69.35	6.64	52.77 – 81.98
True	3	121.02	8.41	95.00 – 133.00	73.83	7.55	57.00 - 87.00
BB		151.79	36.10	72.38 – 183.71	96.96	26.74	40.14 – 120.60
FS		114.89	6.57	90.48 – 129.71	69.14	4.05	55.05 – 78.48
JO-FT		113.47	4.98	99.91 – 156.21	72.97	7.40	54.49 – 104.06
INC-FT		118.35	6.41	90.65 – 126.95	70.64	6.57	58.65 – 81.51
True	4	113.58	6.25	98.00 – 123.00	71.63	10.55	56.00 – 87.00
BB		141.72	41.73	71.45 – 183.40	89.89	31.13	40.21 – 124.37
FS		113.98	6.12	75.25 – 128.27	68.12	5.71	45.23 – 81.55
JO-FT		112.89	4.51	91.57 – 142.95	69.54	6.18	53.10 – 101.39
INC-FT		115.25	5.96	101.40 – 127.83	70.13	8.16	56.00 – 89.06

TABLE VI
COMPARATIVE ANALYSIS WITH THE AAMI STANDARD FOR THE BB, FS, JO-FT, AND INC-FT MODELS BASED ON ST-RESNET OVER THE FOUR-WEEK PERIOD

Model	Week	SBP (mmHg)			DBP (mmHg)		
		ME	S.D.	Conformity	ME	S.D.	Conformity
BB	1	30.96	40.21	No	23.66	32.34	No
	2	28.20	46.19	No	19.69	31.79	No
	3	30.43	38.80	No	23.03	29.02	No
	4	28.31	42.12	No	19.84	30.60	No
	Avg.	29.30	42.31	No	21.30	30.96	No
FS	1	-5.59	6.95	No	-1.97	9.05	No
	2	-10.02	8.89	No	-3.27	7.01	**Yes**
	3	-6.26	6.82	No	-4.75	6.66	**Yes**
	4	0.56	8.51	No	-3.43	10.21	No
	Avg.	-5.36	8.96	No	-3.45	8.37	No
JO-FT	1	-3.08	4.07	**Yes**	0.67	6.20	**Yes**
	2	-1.08	8.62	No	-2.11	8.53	No
	3	-7.89	7.03	No	-3.07	6.56	**Yes**
	4	-0.51	6.25	**Yes**	-3.07	6.56	**Yes**
	Avg.	-3.02	7.55	**Yes**	-1.42	7.84	**Yes**
INC-FT	1	-3.08	4.07	**Yes**	0.67	6.20	**Yes**
	2	-7.54	5.98	No	-1.60	3.28	**Yes**
	3	-2.68	5.89	**Yes**	-3.19	4.76	**Yes**
	4	1.67	5.02	**Yes**	-1.50	9.46	No
	Avg.	-3.00	6.44	**Yes**	-1.56	6.48	**Yes**
AAMI		≤ 5	≤ 8	–	≤ 5	≤ 8	–

weekly results for the FS model were less than those for Grade C for SBP and Grade C for DBP. The average weekly results for the JO-FT model were Grade C for SBP and Grade B for DBP. The results for the INC-FT model were Grade B for both SBP and DBP.

IV. DISCUSSION

A. Performances Between Pretrained, Full-Scratch, and Fine-Tuned Models

As the applicability of deep-learning models for BP prediction from PPG data to other data and the frequency of fine-tuning required when using such models over the long term have been unclear, we assessed the performance of a deep-learning-based BP estimation model that was pretrained on a large-scale dataset or trained on a small-scale dataset from full scratch and explored its adaptability and adjustment frequency during fine-tuning with wearable sensed data over a one-month period. We revealed the effectiveness and adaptability of pretrained and fine-tuned BP estimation models for long-term wearable sensing data. We showed that tuning the models regularly resulted in greater performance than did tuning only once on the first day.

Many deep learning models have been developed for the accurate estimation of BP from PPG data (sometimes with other modalities), and high-performance estimation models based on large datasets have been proposed, with an increasing number of models meeting the standards for evaluating BP estimation [10], [11], [24]. Yen et al. [11] proposed a two-scale long-term recurrent convolutional network with PPGs and electrocardiograms (ECGs) that showed low MAEs (i.e., 3.46 mmHg for SBP and 3.65 mmHg for DBP), meeting the AAMI standard and Grade As for both SBP and DBP

TABLE VII
COMPARATIVE ANALYSIS OF THE BB, FS, JO-FT, AND INC-FT MODELS BASED ON ST-RESNET OVER THE FOUR-WEEK PERIOD ACCORDING TO THE BHS STANDARD

		Cumulative error percentage							
		SBP (mmHg)				DBP (mmHg)			
Model	Week	≤ 5	≤ 10	≤ 15	Grade	≤ 5	≤ 10	≤ 15	Grade
BB	1	0.00	0.00	0.00	–	0.00	0.00	1.11	–
	2	1.31	2.50	4.99	–	0.00	0.00	0.66	–
	3	0.16	0.16	0.32	–	0.16	0.32	0.32	–
	4	0.79	1.26	2.36	–	0.47	1.26	14.78	–
	Avg.	0.66	1.15	2.22	–	0.12	0.37	4.31	–
FS	1	20.18	39.69	71.18	–	41.46	61.42	80.49	–
	2	26.33	39.69	66.02	–	50.71	75.62	99.22	B
	3	44.02	64.11	94.74	–	49.92	76.56	97.13	C
	4	40.47	79.01	91.30	C	36.74	61.74	80.52	–
	Avg.	31.90	61.36	83.29	–	43.24	68.13	91.74	C
JO-FT	1	82.71	96.01	98.44	A	78.05	92.68	98.45	A
	2	15.64	62.29	85.94	–	57.29	77.92	88.70	C
	3	41.71	57.17	92.91	–	65.86	89.37	94.69	B
	4	38.36	65.57	78.30	–	32.23	64.47	74.69	–
	Avg.	42.24	79.83	95.40	C	59.12	81.80	92.69	B
INC-FT	1	82.71	96.01	98.44	A	78.05	92.68	98.45	A
	2	37.09	67.32	80.03	–	69.91	99.48	99.87	A
	3	58.35	72.50	99.36	C	66.14	96.82	99.21	A
	4	67.97	84.76	100.00	B	31.61	65.31	90.49	–
	Avg.	59.10	83.43	93.88	B	59.57	87.37	96.06	B
BHS		≥ 60	≥ 85	≥ 95	A	≥ 60	≥ 85	≥ 95	A
		≥ 50	≥ 75	≥ 90	B	≥ 50	≥ 75	≥ 90	B
		≥ 40	≥ 65	≥ 85	C	≥ 40	≥ 65	≥ 85	C

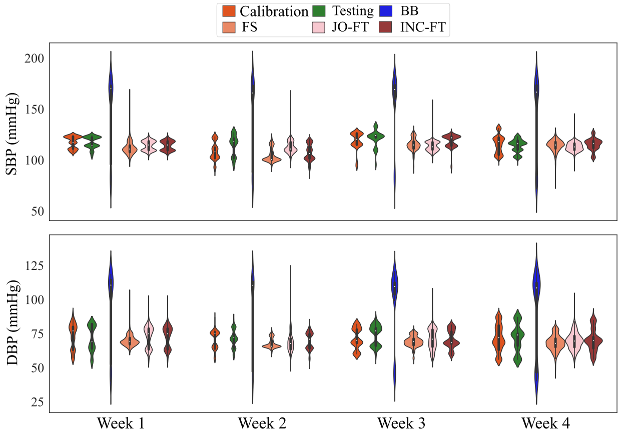


Fig. 8. Violin plots of true and estimated SBP and DBP values from the BB, FS, JO-FT, and INC-FT models based on ST-ResNet over a four-week period.

according to the BHS standard. Zhang et al. [10] retrained the final layer of a pretrained bidirectional long short-term memory (Bi-LSTM)-attention neural network with PPGs from a small number of target subjects in the MIMIC-III dataset for transfer learning. The MAEs were 2.82 mmHg for SBP and 1.88 mmHg for DBP, achieving the AAMI standard and Grade As for both SBP and DBP of the BHS standard. Kanoga et al. [24] reported that the MAEs of an ST-ResNet model based on PPG data were 4.88 mmHg for SBP and 2.60 mmHg for DBP using the MIMIC-III dataset, meeting the AAMI standard and achieving Grade B for SBP and Grade A for DBP according to the BHS standard. They have shown very accurate estimates using deep-learning models when validated on the same dataset.

Several studies have validated PPG-based BP estimation performance of deep learning models via fine-tuning [13], [14], [15], [16] or incremental and continuous learning [17]. The BP estimation performance was improved by splitting a dataset into source and target subsets or by treating different datasets as source and target subsets and by renewing the model entirely or partially by fine-tuning it to correspond to the target subset [13], [14], [15], [16]. In particular, Meng et al. [15] pretrained the MIMIC-III dataset consisting of 50 participants based on an architecture containing convolutional, Bi-LSTM, and dense layers and then used it on a self-collected dataset consisting of 20 participants. The cross-dataset analysis was performed by fine-tuning only the parameters of the final layers of the convolutional and dense layers with 5 min of the self-collected dataset, demonstrating the usefulness of fine-tuning (the MAEs were 3.21 mmHg for SBP and 0.919 mmHg for DBP) [15]. However, fine-tuning was conducted only once, and there was a lack of incremental and continuous learning. Wang et al. [17] measured BPs and PPGs for more than three months and continuously updated the parameters of an error feedback incremental support vector regression and reported that the MAEs of the model were 3.11 mmHg for SBP and 2.47 mmHg for DBP. This study addressed the challenge of training the model with a small initial dataset and gradually adding data to update the model. However, it did not pretrain a model with a large-scale dataset. Thus, the effectiveness and adaptability of pretrained and fine-tuned BP estimation models for long-term wearable sensing data are unclear.

As indicated by the results in Section III, the performance of the model pretrained with the MIMIC-III dataset (i.e., BB model) was poor when using the private dataset obtained from the customized wearable wristband-type PPG device, showing that neither SBP nor DBP met the criteria of the AAMI and BHS standards for all weeks. The BHS standard sets Grade A as a requirement for introduction into medical scenarios. However, the BB model did not meet Grade C. The FS model was much more accurate than the BB model because 6 min of first-session data were used to fully scratch

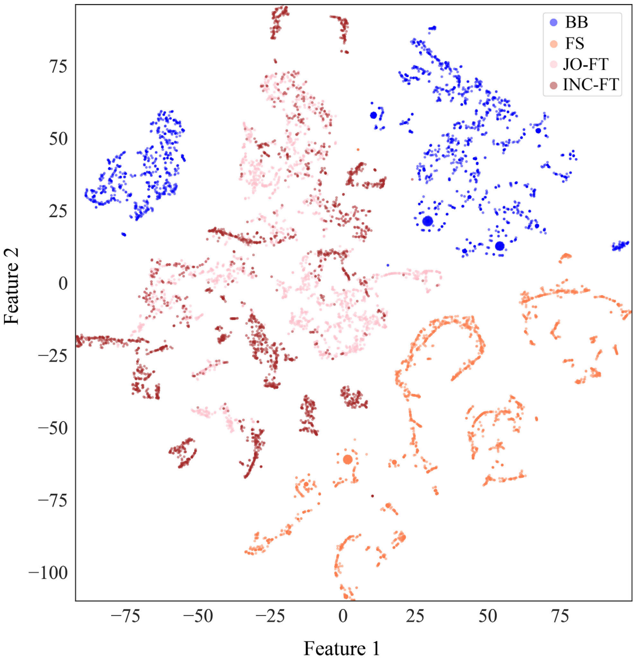


Fig. 9. BB, FS, JO-FT, and INC-FT models for 4-week summary t-SNE mapping.

the model. However, the FS model could not meet the AAMI standard because of the small amount of data compared to the complexity of the model, and in the BHS standard, the SBP was below Grade C, and the DBP was only Grade C. The performances of the fine-tuned (i.e., JO-FT and INC-FT) models were better than those of the BB and FS models. These results indicated that fine-tuning of the BB model is more effective than using the original BB model directly on different datasets or rather than simply constructing an FS model with a small amount of first-session data. In addition, we found no significant differences in the estimation performance between the JO-FT and INC-FT models with regard to DBP estimation (see Table IV). Fine-tuning may not need to be performed every week in some cases, although the performance is better if fine-tuning is performed every week and not only on the first day. However, even the most accurate fine-tuned model, the INC-FT model, did not simultaneously meet the AAMI and BHS standards for all weeks and has not reached the level at which it can be used for medical applications. Therefore, improving the estimation performance remains a challenge.

To show the feasibility of estimating BP from fine-tuned models, we used the t-distributed stochastic neighbor embedding (t-SNE) method, which nonlinearly reduces the original dimension to a lower dimension in an unsupervised manner and preserves the clustering relationships (Fig. 9). The clusters are independent in the BB and FS models but are similarly scattered between the JO-FT and INC-FT models. This may be because the feature space obtained from the MIMIC-III dataset, which is the source, is being shifted to the feature space of the private dataset, which is the target. In addition, the INC-FT model incorporates features from multiple days, resulting in a larger variance. Therefore, BP estimation with the INC-FT model may have been more accurate because it was able to use a good combination of sources and target features that successfully incorporated data from multiple days.

B. Long-Term Use of BP Estimators and Fine-Tuning Approaches

It is important to achieve a high overall performance for routine daily BP determination; however, BP estimation modules that are likely to be used over a long period, such as health monitoring, need to be able to provide stable and high performance. Several studies have reported on the long-term use of BP estimators. Su et al. [28] investigated the long-term follow-up performance of deep recurrent neural network models using PPGs and ECGs for estimating BPs at the first, second, and fourth days and again six months later by fine-tuning the final layer of the model that was pretrained in 84 participants with some data obtained from 12 participants on the first day. They found that the model showed low root mean square errors (1.80–5.21 mmHg for SBP and DBP of 3.84–5.81 mmHg for DBP from the first day to six months later). Yao et al. [29] trained artificial neural networks with data obtained from 33 individuals with an integrated multidimensional feature set of wrist PPGs as input and tested their long-term estimation performance from 3 to 15 days for two individuals. Liu et al. [30] evaluated several machine learning and deep learning models established using large datasets of PPGs and ECGs obtained from 3077 individuals via smartwatches and tracked their long-term accuracy. Their models were calibrated with individualized first-day BP values every other week over a one-month period and showed an increase in error [30].

Although BP estimation over a long-term period has been performed in the abovementioned studies [28], [29], [30], our study showed the effectiveness of fine-tuning a BP estimator, which was pretrained on a public dataset, for application to other long-term datasets. The BB, FS, JO-FT, and INC-FT models were assessed using PPG and cuff-based BP data obtained from 11 subjects on four days in a month. The amount of PPG data obtained per person per day was 12 min. As shown in Fig. 7, the BP values did not follow a simple normal distribution and varied slightly from week to week. If training and test data have the same distribution, this complex representation can be acquired by a deep learning model, and the test performance will be high. However, if the data distribution fluctuates, as in this study, it is better to fine-tune the estimated BP for each week, even if this is done only once, to obtain the true value. In this study, this approach yielded a significantly closer estimate (see Fig. 8). As seen from the BPs estimated by using INC-FT, the estimated BPs were still closer to the true distribution when fine-tuning was carried out occasionally by using the calibration data for the relevant week. Measuring the calibration data daily would be ideal for maintaining the model's performance. Nevertheless, such cumbersome requirements reduce the usefulness of the application. Reconciliation of the trade-off between convenience and performance remains a practical challenge in this field.

Another interesting point is that the public and private datasets were successfully tuned despite different measurement locations, devices, and subject groups. The public dataset, which included data obtained from individuals' fingertips using commercial pulse oximetry devices, and the private dataset, obtained from individuals' wrists by means of a customized wearable wristband-type device, involved different measurement locations and devices. In addition, the subject groups differed significantly between ICU patients and healthy subjects. However, over a longer period (i.e., one month), fine-tuning significantly improved the BP estimation. If a model can be pretrained with large-scale datasets to obtain a variety of representations and use a small amount of calibration data, it would be possible to construct a BP estimator that meets the AAMI and BHS standards.

C. Limitations

To achieve robust BP estimation in daily life while meeting the AAMI and BHS standards, the number of participants needs to be increased, and the measurement days need to be more closely spaced. This study measured PPG and BP data from 11 healthy participants over a one-month period to evaluate the pretrained and fine-tuned deep learning models. The young, healthy participants included in this study had a narrow distribution of BP values; thus, models trained using these data cannot be generalized to a wider range of BP distributions or age groups [30]. In other words, the performance of the fine-tuning model shown in the present study may not be directly applicable to older or unhealthy people (e.g., hypertensive patients) who were not included in the private dataset. In addition, the number of measurement days was limited because this dataset included only four days per month.

However, data measured over a long period, as in this study, are valuable, as our findings may promote further studies in this field. Therefore, we have made this dataset publicly available at: https://github.com/aistairc/One_Month_Wrist_PPG_Dataset.

V. Conclusion

This study sheds light on the effectiveness and adaptability of pretrained and fine-tuned BP estimation models when using long-term wearable sensor data. We investigated the performance of deep-learning BP estimation models that were pretrained using a large-scale dataset (MIMIC-III) and then fine-tuned with data from a smaller cohort obtained via a wearable wristband-type PPG sensor over one month.

Our findings demonstrated the superior performance of the fine-tuned models compared with their non-fine-tuned counterparts. The daily fine-tuning of the model yielded greater performance than did the fine-tuning of the model only once on the first day. This delicate balance between convenience and performance presents practical challenges. Despite the accuracy of our most finely tuned model, it consistently falls short of meeting medical standards, posing a significant obstacle to medical application. Enhancing the estimation performance remains a challenge in this field.

Institutional Review

The study was conducted in accordance with the guidelines of the Declaration of Helsinki and approved by the Institutional Review Board of the National Institute of Advanced Industrial Science and Technology (AIST), Japan (protocol code: HF2022-1204; date of approval: May 2, 2022).

Informed Consent

Informed consent was obtained from all subjects involved in the study.

Data Availability

The public dataset used in this study is available from PhysioNet https://physionet.org/content/mimic3wdb-matched/1.0/ (accessed October 10, 2023) [19]. A preprocessed public dataset is available at https://drive.google.com/drive/folders/182xF0Y8NPiDGownNcxUfE4D-BPwv-bvO?usp=drive_link (accessed on October 10, 2023) [24].

Acknowledgment

The authors would like to thank Yuko Kawai and Rika Yoshimuta for helping with the data measurements.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Z. N. Hasan, M. Q. Hussein, and G. F. Haji, "Hypertension as a risk factor: Is it different in ischemic stroke and acute myocardial infarction comparative cross-sectional study?" *Int. J. Hypertension*, vol. 2011, pp. 1–5, Oct. 2011.

[2] *WHO Reveals Leading Causes of Death and Disability Worldwide: 2000–2019*, World Health Org., Geneva, Switzerland, 2020.

[3] B. Zhou et al., "Worldwide trends in hypertension prevalence and progress in treatment and control from 1990 to 2019: A pooled analysis of 1201 population-representative studies with 104 million participants," *Lancet*, vol. 398, no. 10304, pp. 957–980, 2021.

[4] P. K. Whelton et al., "2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA guideline for the prevention, detection, evaluation, and management of high blood pressure in adults: A report of the American college of cardiology/American heart association task force on clinical practice guidelines," *Hypertension*, vol. 71, no. 6, pp. e127–e248, Jun. 2018.

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

[6] J. Esmaelpoor, M. H. Moradi, and A. Kadkhodamohammadi, "A multistage deep neural network model for blood pressure estimation using photoplethysmogram signals," *Comput. Biol. Med.*, vol. 120, May 2020, Art. no. 103719.

[7] M. Panwar, A. Gautam, D. Biswas, and A. Acharyya, "PP-Net: A deep learning framework for PPG-based blood pressure and heart rate estimation," *IEEE Sensors J.*, vol. 20, no. 17, pp. 10000–10011, Sep. 2020.

[8] D. Wang, X. Yang, X. Liu, L. Ma, L. Li, and W. Wang, "Photoplethysmography-based blood pressure estimation combining filter-wrapper collaborated feature selection with LASSO-LSTM model," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–14, 2021.

[9] C.-T. Yen, J.-X. Liao, and Y.-K. Huang, "Applying a deep learning network in continuous physiological parameter estimation based on photoplethysmography sensor signals," *IEEE Sensors J.*, vol. 22, no. 1, pp. 385–392, Jan. 2022.

[10] Y. Zhang, X. Ren, X. Liang, X. Ye, and C. Zhou, "A refined blood pressure estimation model based on single channel photoplethysmography," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 12, pp. 5907–5917, Dec. 2022.

[11] C.-T. Yen, S.-N. Chang, and C.-H. Liao, "Estimation of beat-by-beat blood pressure and heart rate from ECG and PPG using a fine-tuned deep CNN model," *IEEE Access*, vol. 10, pp. 85459–85469, 2022.

[12] A. Kumar, R. Komaragiri, and M. Kumar, "Blood pressure estimation and classification using a reference signal-less photoplethysmography signal: A deep learning framework," *Phys. Eng. Sci. Med.*, vol. 46, no. 4, pp. 1589–1605, Dec. 2023.

[13] W. Wang, P. Mohseni, K. L. Kilgore, and L. Najafizadeh, "Cuff-less blood pressure estimation from photoplethysmography via visibility graph and transfer learning," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 5, pp. 2075–2085, May 2022.

[14] J. Leitner, P.-H. Chiang, and S. Dey, "Personalized blood pressure estimation using photoplethysmography: A transfer learning approach," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 1, pp. 218–228, Jan. 2022.

[15] Z. Meng, X. Yang, X. Liu, D. Wang, and X. Han, "Non-invasive blood pressure estimation combining deep neural networks with pre-training and partial fine-tuning," *Physiological Meas.*, vol. 43, no. 11, Nov. 2022, Art. no. 11NT01.

[16] Z. Liu, Y. Zhang, and C. Zhou, "BiGRU-attention for continuous blood pressure trends estimation through single channel PPG," *Comput. Biol. Med.*, vol. 168, Jan. 2024, Art. no. 107795.

[17] D. Wang, X. Yang, J. Wu, and W. Wang, "Personalized modeling of blood pressure with photoplethysmography: An error-feedback incremental support vector regression model," *IEEE Internet Things J.*, vol. 11, no. 1, pp. 1732–1745, Jul. 2023.

[18] D. Liu, M. Görges, and S. A. Jenkins, "University of Queensland vital signs dataset: Development of an accessible repository of anesthesia patient monitoring data for research," *Anesthesia Analgesia*, vol. 114, no. 3, pp. 584–589, Mar. 2012.

[19] A. E. W. Johnson et al., "MIMIC-III, a freely accessible critical care database," *Sci. Data*, vol. 3, no. 1, pp. 1–9, May 2016.

[20] V. Hartmann, H. Liu, F. Chen, Q. Qiu, S. Hughes, and D. Zheng, "Quantitative comparison of photoplethysmographic waveform characteristics: Effect of measurement site," *Frontiers Physiol.*, vol. 10, pp. 1–19, Mar. 2019.

[21] M.-T. Wu, I.-F. Liu, Y.-H. Tzeng, and L. Wang, "Modified photoplethysmography signal processing and analysis procedure for obtaining reliable stiffness index reflecting arteriosclerosis severity," *Physiological Meas.*, vol. 43, no. 8, Aug. 2022, Art. no. 085001.

[22] *American National Standards for Electronic or Automated Sphygmomanometers*, Standard ANSI/AAMI SP 10-1987, Assoc. for Advancement Med. Instrum., 1987.

[23] E. O'Brien et al., "The British hypertension society protocol for the evaluation of automated and semi-automated blood pressure measuring devices with special reference to ambulatory systems," *J. Hypertension*, vol. 8, no. 7, pp. 607–619, Jul. 1990.

[24] S. Kanoga et al., "Comparison of seven shallow and deep regressors in continuous blood pressure and heart rate estimation using single-channel photoplethysmograms under three evaluation cases," *Biomed. Signal Process. Control*, vol. 85, Aug. 2023, Art. no. 105029.

[25] Q. Hu, X. Deng, A. Wang, and C. Yang, "A novel method for continuous blood pressure estimation based on a single-channel photoplethysmogram signal," *Physiological Meas.*, vol. 41, no. 12, Dec. 2020, Art. no. 125009.

[26] M. Kachuee, M. M. Kiani, H. Mohammadzade, and M. Shabany, "Cuffless blood pressure estimation algorithms for continuous health-care monitoring," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 4, pp. 859–869, Apr. 2017.

[27] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 1–6.

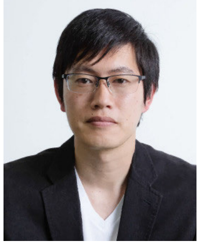
[28] P. Su, X.-R. Ding, Y.-T. Zhang, J. Liu, F. Miao, and N. Zhao, "Long-term blood pressure prediction with deep recurrent neural networks," in *Proc. IEEE EMBS Int. Conf. Biomed. Health Informat. (BHI)*, Mar. 2018, pp. 323–328.

[29] P. Yao et al., "Multi-dimensional feature combination method for continuous blood pressure measurement based on wrist PPG sensor," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 8, pp. 3708–3719, Aug. 2022.

[30] Z.-D. Liu et al., "Cuffless blood pressure measurement using smartwatches: A large-scale validation study," *IEEE J. Biomed. Health Informat.*, vol. 27, no. 9, pp. 1–13, Sep. 2023. [Online]. Available: https://ieeexplore.ieee.org/document/10129795

Satoshi Kamei received the B.S. degree from Tokyo University of Science, Chiba, Japan, in 2018, where he is currently pursuing the degree in engineering.

His research interests include machine learning, implicit sensing, and utilization of biological information.



Suguru Kanoga received the M.E. and Ph.D. degrees in engineering from Keio University, Kanagawa, Japan, in 2014, and 2016, respectively.

He is currently a Senior Researcher at the Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology (AIST), Tokyo, Japan. His research interests include the development of machine learning techniques, particularly robust feature extraction and implementation of human–computer interfaces based on the proposed machine learning techniques.

Dr. Kanoga was a Research Fellow of the Japan Society for the Promotion of Science, Japan, from 2015 to 2017.

Masataka Yamamoto (Member, IEEE) received the M.S. degree in human science from Nihon University, Tokyo, Japan in 2015, and the Ph.D. degree in engineering from Hiroshima University, Hiroshima, Japan, in 2018.

In 2018, he joined Hiroshima University, as a Researcher. Since 2019, he has been an Assistant Professor at Tokyo University of Science, Chiba, Japan, and a Visiting Lecturer at Hiroshima University. His research interests include biomechanics, rehabilitation, gait analysis, and human modeling.

Hiroshi Takemura (Member, IEEE) received the M.S. degree from the Kyushu Institute of Technology, Kitakyushu, Japan, in 2001, and the Ph.D. degree in information engineering from the Nara Institute of Science and Technology, Ikoma, Japan, in 2003.

From October 2003 to March 2004 and from January 2005 to May 2005, he was employed as a Lecturer (PT) at Nara Institute of Science and Technology, Nara, Japan. From April 2004 to December 2004, he attended the University of Karlsruhe, Karlsruhe, Germany, as a Guest Lecturer at Industrial Applications of Information and Microsystems. In 2005, he joined the Department of Mechanical Engineering, Tokyo University of Science, Chiba, Japan, where he has been a Professor since 2019. His research interests include robotics, medical devices, biomechanics, gait analysis, and human modeling.

Mitsunori Tada received the Ph.D. degree in engineering from Nara Institute of Science and Technology, Nara, Japan, in 2002.

Since 2002, he has been at the National Institute of Advanced Industrial Science and Technology (AIST), Tokyo, Japan, where he has been the Leader of the Research Team, Artificial Intelligence Research Center, since 2018. His research interests include real-time human motion measurement and analysis to realize human-centered cyber–physical systems with human digital twins.