Cuffless Blood Pressure Measurement Using Smartwatches: A Large-Scale Validation Study

Zeng-Ding Liu, Ye Li, *Senior Member, IEEE*, Yuan-Ting Zhang, *Fellow, IEEE*, Jia Zeng, Zu-Xian Chen, Zhi-Wei Cui, Ji-Kui Liu, and Fen Miao, *Member, IEEE*

***Abstract*—This study aimed to evaluate the performance of cuffless blood pressure (BP) measurement techniques in a large and diverse cohort of participants. We enrolled 3077 participants (aged 18–75, 65.16% women, 35.91% hypertensive participants) and conducted followed-up for approximately 1 month. Electrocardiogram, pulse pressure wave, and multiwavelength photoplethysmogram signals were simultaneously recorded using smartwatches; dual-observer auscultation systolic BP (SBP) and diastolic BP (DBP) reference measurements were also obtained. Pulse transit time, traditional machine learning (TML), and deep learning (DL) models were evaluated with calibration and calibration-free strategy. TML models were developed using ridge regression, support vector machine, adaptive boosting, and random forest; while DL models using convolutional and recurrent neural networks. The best-performing calibration-based model yielded estimation errors of 1.33 ± 6.43 mmHg for DBP and 2.31 ± 9.57 mmHg for SBP in the overall population, with reduced SBP estimation errors in normotensive (1.97 ± 7.85 mmHg) and young (0.24 ± 6.61 mmHg) subpopulations. The best-performing calibration-free model had estimation errors of −0.29 ± 8.78 mmHg for DBP and −0.71 ± 13.04 mmHg for SBP. We conclude that smartwatches are effective for measuring DBP for all participants and SBP for normotensive and younger participants with calibration; performance degrades significantly for heterogeneous populations including older and hypertensive participants. The availability of cuffless BP measurement without calibration is limited in routine settings. Our study provides a large-scale benchmark for emerging investigations on cuffless BP measurement, highlighting the need to explore additional signals or principles to enhance the accuracy in large-scale heterogeneous populations.**

***Index Terms*—Benchmark, cuffless blood pressure, deep learning, large-scale validation study, machine learning.**

I. Introduction

HYPERTENSION is a major risk factor for cardiovascular disease [1], [2], affecting more than 1 billion adults worldwide [3]. Despite the increasing attention given to the control of hypertension, fewer than half of adults with hypertension are diagnosed and treated appropriately [3]. Regular monitoring of blood pressure (BP) plays a vital role in the early detection of hypertension. However, clinical BP may be inaccurate due to the masked and white coat effects [4]. Ambulatory BP monitoring (ABPM), in which BP is assessed over a 24-hour period to achieve a timely and accurate hypertension diagnosis, has been proven to be superior to clinical measurements in predicting cardiovascular mortality [5], [6]. However, existing ABPM techniques rely on an inflatable cuff, which can disturb the sleep and daily activities of users [7], thus limiting their usage.

Cuffless BP measurement approaches have been proposed to overcome the limitation of cuff-based ABPM techniques [8], enabling unobtrusive and continuous monitoring of BP through wearable devices [9], [10]. Over the past two decades, the development of BP measurement techniques has evolved from mechanism-driven solutions [11], [12], [13], [14], [15] to data-driven solutions [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29]. In mechanism-driven solutions, formulas for estimating BP are developed based on specific indicators reflecting BP changes from known hemodynamic principles and/or autonomic regulation functions. The most popular indicator is arterial pulse transit time (PTT), which is the time it takes an arterial pulse wave to travel from one arterial site to another [30]. Arterial PTT can be obtained as the time span between an electrocardiogram (ECG) signal and a pulse wave or as the time delay between two pulse waves [8].

Manuscript received 29 December 2022; revised 6 April 2023 and 7 May 2023; accepted 11 May 2023. Date of publication 19 May 2023; date of current version 6 September 2023. This work was supported in part by the National Natural Science Foundation of China under Grant U1913210, in part by Strategic Priority CAS Project under Grant XDB38040200, and in part by the Basic Research Project of Shenzhen under Grants (JCYJ20220818101216034 and JCYJ20210324101206017). *(Zeng-Ding Liu and Ye Li contributed equally to this work.) (Corresponding author: Fen Miao.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Ethics Committee of Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences Application No. 210315-H0558.

Zeng-Ding Liu is with the Key Laboratory for Health Informatics, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518172, China, and also with the University of Chinese Academy of Sciences, Beijing 101408, China (e-mail: zd.liu@siat.ac.cn).

Ye Li and Fen Miao are with the Key Laboratory for Health Informatics, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518172, China (e-mail: ye.li@siat.ac.cn; fen.miao@siat.ac.cn).

Yuan-Ting Zhang is with the Hong Kong Center for Cerebro-Caridovascular Health Engineering (COCHE), Hong Kong 999077, SAR, China, and also with the Department of Biomedical Engineering, City University of Hong Kong, Hong Kong 999077, SAR, China (e-mail: yt.zhang@cityu.edu.hk).

Jia Zeng, Zu-Xian Chen, and Zhi-Wei Cui are with the Huawei Technologies Company, Ltd., Shenzhen 518129, China (e-mail: j.zeng@huawei.com; chenzuxian1@huawei.com; cuizhiwei3@huawei.com).

Ji-Kui Liu is with the Key Laboratory for Health Informatics, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518172, China, and also with the Shenzhen Polytechnic, Shenzhen 518055, China (e-mail: jk.liu@siat.ac.cn).

Digital Object Identifier 10.1109/JBHI.2023.3278168

Various pulse waves-based techniques, such as photoplethysmogram (PPG) [11] and pulse pressure wave (PPW) [18], [31] techniques, have been proposed for arterial PTT calculations. Although pioneering mechanism-driven solutions have yielded promising results for cuffless BP measurement, their accuracy varies greatly along with different population characteristics and follow-up duration.

To improve the accuracy of BP estimation, many researchers have shifted their focus from mechanism-driven solutions to data-driven solutions in recent years. In data-driven solutions, the features leading to BP changes are manually defined or automatically learned. Data mining algorithms are then used to construct a mapping between these features and BP. On the basis of the feature extraction method, data-driven solutions can be further categorized into traditional machine learning (TML)-based [16], [17], [18], [19], [20], [21] and deep learning (DL)-based [22], [23], [24], [25], [26], [27], [28], [29] approaches. In TML-based approaches, physiological features associated with BP are manually extracted from raw signals and then translated into BP values using TML algorithms, such as ridge regression [21], support vector machine (SVM) [17], adaptive boosting (AdaBoost) [16], and random forest (RF) [19]. With the advances in end-to-end feature learning techniques in DL, many researchers have shifted their focus from traditional feature-engineering-based approaches to DL-based approaches for BP measurement. Various DL algorithms, such as recurrent neural network [23], [24], convolutional neural network [25], [28], and transfer learning [26], [27], have been used in such approaches.

Despite considerable advances in cuffless BP measurement techniques have been made, their performance has not been fully validated, and thus these techniques are not universally accepted [32], [33]. First, many cuffless BP models were validated in small, young, and healthy populations under controlled experimental settings. Therefore, these models may not be generalizable to large-scale heterogeneous populations. Moreover, some models have performed sub optimally in real-world tests involving patients with hypertension [13], [15]. Second, whether the calibration-based models remain usable over the long-term is the most important concern; however, this issue has not been fully validated. Third, the reference employed in many studies is oscillometric BP or Finapres BP [8]; nevertheless, dual-observer auscultation BP is the gold reference depending on the American National Standards Institute/ Association for the Advancement of Medical Instrumentation/ International Organization for Standardization (ANSI/AAMI/ISO) guidelines [34]. Imprecise reference would result in biased models [21]. Thus, an equitable evaluation platform is necessary for cuffless BP measurements.

In consideration of the aforementioned problems, in this paper, a large-scale cuffless BP dataset named **CAS-BP**[1] is first constructed. A benchmark for evaluating cuffless BP measurement methods is then developed. Our work has the following advantages: 1) This is the largest known validation study for evaluating the performance of smartwatch-based cuffless BP measurement techniques with dual-observer auscultation systolic BP (SBP) and diastolic BP (DBP) as the reference; 2) The protocol and participant classification were designed exactly according to the ANSI/AAMI/ISO standard and thus have good generalizability to real-world settings; 3) Six-channel signals, including ECG, PPW, and multi-wavelength PPG (MWPPG), were simultaneously recorded using smartwatches and evaluated for BP estimation.

II. Related Works

Cuffless BP measurement methods can be generally categorized into mechanism-driven and data-driven solutions. In the following paragraphs, we will review some related works regarding these categories.

A. Mechanism-Driven Approaches

The arterial PTT-based model constitutes the most popular mechanism-driven approach for cuffless BP measurements. According to the Moens-Korteweg and Hughes equations [35], increased arterial stiffness results in faster pulse wave propagation in arteries (i.e., decreased arterial PTT) and increased arterial BP. Therefore, arterial BP is inversely proportional to arterial PTT. Many arterial PTT-based BP estimation models have been developed based on this principle [8]. In 2005, Poon et al. [11] proposed a widely cited arterial PTT-based algorithm for SBP and DBP estimation, as expressed in (1):

$$DBP = MBP_0 + \frac{2}{\gamma}\ln\left(\frac{PTT_0}{PTT}\right) + \frac{PP_0}{3}\cdot\left(\frac{PTT_0}{PTT}\right)^2 \tag{1a}$$

$$SBP = DBP + PP_0\cdot\left(\frac{PTT_0}{PTT}\right)^2 \tag{1b}$$

where PTT is arterial PTT, $MBP_0 = (SBP_0 + 2DBP_0)/3$, $PP_0 = SBP_0 - DBP_0$, and SBP_0, DBP_0, and PTT_0 are measured values that are used for calibration; γ is the subject-dependent coefficient. Experimental results obtained from 85 individuals indicated that the algorithm performs well in BP measurement according to the ANSI/AAMI/ISO standard [34]. In 2016, Ding et al. [12] reported that PPG intensity ratio (PIR) can capture variations in arterial diameter, enabling tracking low frequency BP, i.e., DBP. They proposed a BP model that fuses arterial PTT and PIR, as expressed in (2):

$$DBP = DBP_0\cdot\frac{PIR_0}{PIR} \tag{2a}$$

$$SBP = DBP + PP_0\cdot\left(\frac{PTT_0}{PTT}\right)^2 \tag{2b}$$

where DBP_0, $PP_0 = SBP_0 - DBP_0$, PTT_0, and PIR_0 are measured values for calibration. Experimental results for 27 healthy individuals revealed that their proposed method outperformed conventional arterial PTT algorithms, achieving estimation errors of −0.37 ± 5.21 mmHg for SBP and −0.18 ± 4.13 mmHg for DBP. Furthermore, because arteriolar PTT calculated from MWPPG can be used as an indicator of systemic vascular resistance, Liu et al. [13] proposed an arteriolar PTT-based model for BP measurement, as described in (3):

$$MBP = HR\cdot(k_1\cdot ePTT + k_2) \tag{3a}$$

$$PP = MBP\cdot\left(k_2\cdot\frac{t_\tau}{HR} + b_2\right) \tag{3b}$$

where MBP and PP are mean BP and pulse pressure, respectively, which can be converted into SBP and DBP by $MBP = (SBP + 2DBP)/3$ and $PP = SBP - DBP$; k_1, b_1, k_2, and b_2 are individualized parameters and can be obtained

[1] CAS-BP dataset is available at [Online]. Available: https://github.com/zdzdliu/CAS-BP.

by least squares regression; HR is heart rate; $ePTT$ is arteriolar PTT; t_τ is a time constant and can be extracted from a PPG waveform [13]. Experimental results for 20 individuals demonstrated that the proposed arteriolar PTT-based model has better accuracy in tracking BP than the conventional arterial PTT-based method.

Although PTT-based methods have been gradually refined by introducing new physiological features, they still typically have low accuracy and robustness [8] primarily because PTT-based methods are based on a fixed hypothesis; and only include a small part of factors affecting BP. However, the factors leading to the variation of BP is complex in real-world settings, including cardiac output, vascular tone, and physiological status.

B. Data-Driven Approaches

Data-driven approaches can achieve greater BP estimation accuracy than mechanism-driven approaches by applying TML or DL algorithms to automatically construct the complex relationships between physiological signals and BP values. For TML-based BP estimation, meaningful handcrafted features are extracted to develop the model. In addition to PTT, pulse morphological features calculated from a pulse wave (PPG or PPW) waveform and its derivatives have been proven useful for BP estimation [36]. Thus, various pulse wave features, such as time-, slope-, and area-related features, have been proposed in the BP estimation literature [16], [17], [18], [19], [20], [21].

Miao et al. [18] estimated SBP and DBP by inputting 35 features extracted from ECG and PPW signals to a multi-instance regression algorithm, achieving estimation errors of 1.62 ± 7.76 mmHg for SBP and 1.49 ± 5.52 mmHg for DBP on 85 individuals. Yang et al. [19] utilized 42 features extracted from ECG and PPG signals to estimate SBP and DBP with various TML algorithms, including linear regression, RF, artificial neural network (ANN), and recurrent neural network. Their lowest estimation errors for SBP and DBP in a database of surgical patients were 0.05 ± 6.92 mmHg and -0.05 ± 3.99 mmHg, respectively. These methods required two-channel signals. However, to reduce the sensing burden of BP devices, several studies have developed BP models based on one-channel signals. Haddad et al. [20] presented a linear regression model to estimate BP by using 27 features that were calculated only from PPG and its derivatives and achieved favorable accuracy on the public Medical Information Mart for Intensive Care (MIMIC) database. Yao et al. [37] proposed a multi-dimensional feature combination method based on basis demographics [age, height, weight, body mass index (BMI), and gender] and three groups (time-domain, morphological, and statistical) of PPG features input to an ANN algorithm. Their constructed model was examined on a dataset of 33 individuals and achieved good accuracy for both SBP and DBP estimation. Recently, Microsoft Research team evaluated the performance of tonometry, PPG, and ECG signals for estimating BP by using ridge regression on a relatively large population (1125 participants) [21], [33]. Their findings suggested that tonometry-derived features were superior to other features calculated from PPG and ECG for estimating BP. However, the performance of all these methods is strongly affected by signal preprocessing process (i.e., pinpointing the location of feature points in the signal) due to the need to calculate handcrafted features, especially in real-world setting with strong noise.

In recent years, DL-based cuffless BP estimation approaches have attracted the interest of researchers [22], [23], [24], [25], [26], [27], [28], [29]. These approaches automatically learn representative features from raw signals, avoiding handcrafted feature extraction. Liu et al. [22] verified the possibility to estimate BP from PPW signals with the VGGNet architecture on a dataset of 89 individuals. Fan et al. [24] presented a bidirectional long short-term memory network (BiLSTM) to estimate BP values from one-channel ECG signals, achieving estimation errors of 0.18 ± 10.83 for SBP and 1.24 ± 5.90 mmHg for DBP on the MIMIC database. Similarly, Miao et al. [25] proposed a hybrid network that fused a residual network and long short-term memory (ResLSTM) for cuffless BP estimation using only ECG signals. Kim et al. [28] proposed a DL architecture combining self-attention and U-Net for estimating BP from PPG signals, reporting estimation errors of 1.23 ± 5.40 mmHg for SBP and -0.53 ± 2.81 mmHg for DBP on the MIMIC database. Wang et al. [26] introduced a transfer learning approach for cuffless BP measurement based on short-duration PPG signals. They created images from PPG signals using visibility graphs and applied pretrained deep convolutional neural networks to extract features from these images to estimate BP. In experiments on the MIMIC database, the proposed method yielded estimation errors of 0.00 ± 8.46 mmHg and -0.04 ± 5.36 mmHg for SBP and DBP, respectively. Although these studies had favorable results on the MIMIC database, it is worth noting that the database was acquired in a particular setting (i.e., intensive care units) with physiological signals collected by medical instruments. Hence, it is unclear whether these results can be replicated in large-scale heterogeneous populations in routine settings using wearables.

III. Materials and Methods

Fig. 1 presents the framework diagram of the BP model construction in this study. The steps comprise data collection, data preprocessing, and the construction and evaluation of BP model. The details are described in the following sections.

A. Experimental Protocol

This study recruited 3077 individuals without severe cardiovascular diseases or behavioral disorders to participate in a follow-up experiment lasting approximately 1 month. ECG, MWPPG (four-channel PPG with varying wavelengths), and PPW signals were simultaneously acquired by a smartwatch. The smartwatch was a prototype supplied by Huawei Technologies, equipped with an ECG sensor, a PPW sensor, and an MWPPG sensor. Fig. 2(a) (right) shows the placement of these sensors to obtain the signals. Specifically, two ECG electrodes located on the back of the dial should be attached to the left wrist, while the third electrode pressed by the participant's right thumb. The MWPPG sensor should be pressed by the participant's right index finger. The PPW sensor, fitted with piezoelectric materials on the strap, should be placed at the radial artery to measure PPW signal. A detailed description of the PPW measurement principle can be found in our previous study [18], [31]. Note that the MWPPG sensor comprised four LEDs with wavelengths of 940 nm (infrared), 650 nm (red), 590 nm (yellow), and 470 nm (blue); the four channels were thus denoted PPGIR, PPGR, PPGY, and PPGB, respectively. Corresponding reference BP values were measured using a cuff-based, clinically validated dual-stethoscope mercury sphygmomanometer (Yuwell YE670AH, Yuwell Medical Equipment Co, China).

The experimental procedure was carried out in the following steps (Fig. 2(b)). i) *Preparation:* The participant was asked to sit quietly for 5 minutes before the measurement. ii) *First*

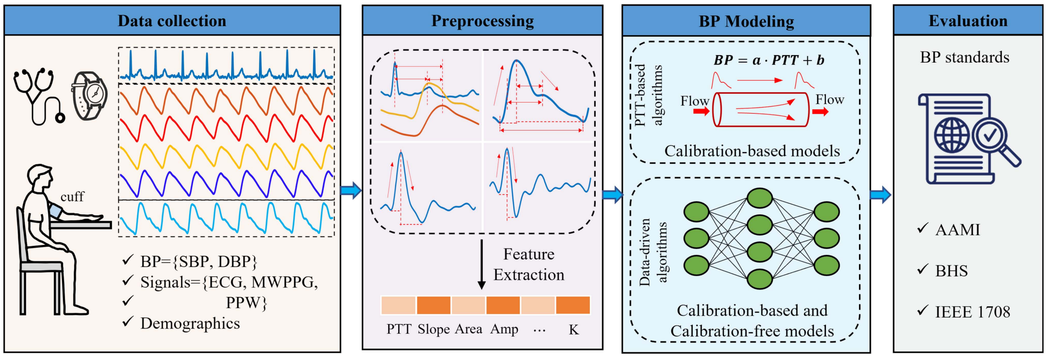

Fig. 1. Block diagram of BP modeling.

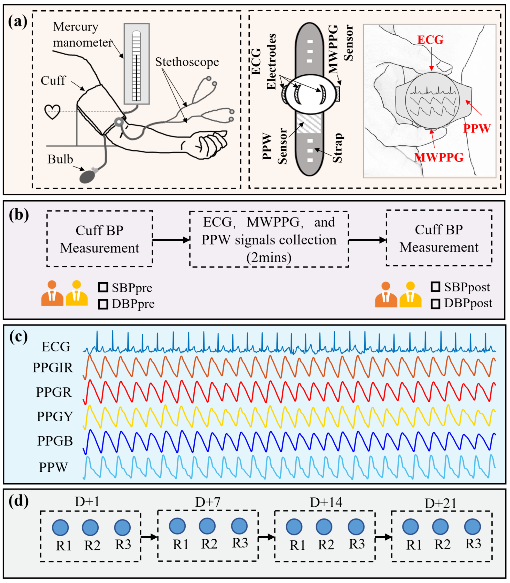

Fig. 2. Experimental protocol. (a) Dual-stethoscope mercury sphygmomanometer for cuff BP measurement (left) and smartwatch for signal collection (right). (b) Recording collection process. (c) An example of the collected signals. The MWPPG signals comprised four channels: PPGIR, PPGR, PPGY, and PPGB, acquired from infrared, red, yellow, and blue light, respectively. (d) Each participant underwent a total of 12 recordings over four different days: D, D+7, D+14, and D+21, with three recordings (R1, R2, and R3) taken per day.

TABLE I
PARTICIPANTS CHARACTERISTIC IN **CAS-BP** DATASET

Characteristics	CAS-BP Dataset	ANSI/AAMI/ISO Requirement
Subjects, n	3077	$\geq$ 85
Age, years	46.32 ± 16.92	$>$ 12
Male, %	34.84	$\geq$ 30
Female, %	65.16	$\geq$ 30
BMI, kg/m^2	23.84 ± 3.4 3	-
History of hypertension, %	35.91	-
BP recordings, n	29568	$\geq$ 255
SBP, mmHg	123.6 ± 21.2	-
DBP, mmHg	83.4 ± 12.1	-
SBP $\geq$ 160 mmHg, %	5.53	$\geq$ 5
SBP $\geq$ 140 mmHg, %	24.82	$\geq$ 20
SBP $\leq$ 100 mmHg, %	12.77	$\geq$ 5
DBP $\geq$ 100 mmHg, %	4.27	$\geq$ 5
DBP $\geq$ 85 mmHg, %	21.76	$\geq$ 20
DBP $\leq$ 60 mmHg, %	5.33	$\geq$ 5

BP measurement: Two trained and mutually blinded observers measured auscultation SBP and DBP from the left upper arm using a standardized procedure [38]. If the difference in the BP (SBP or DBP) values between the two observers was $\leq$ 5 mmHg, the average value was used as the reference value. Otherwise, the measurement was repeated. The SBP and DBP measured in this step were denoted as $\mathrm{SBP}_{\mathrm{pre}}$ and $\mathrm{DBP}_{\mathrm{pre}}$, respectively. iii) *Signal acquisition:* After the BP measurement, the participant wore a smartwatch on the left wrist for 2 minutes to simultaneously acquire the ECG, radial PPW, and finger MWPPG signals (Fig. 2(c)). The sampling frequency was 200 Hz for PPW signals and 1000 Hz for ECG and MWPPG signals. iv) *Second BP measurement:* After measuring the signals, repeated step ii), and the measured SBP and DBP were denoted as $\mathrm{SBP}_{\mathrm{post}}$ and $\mathrm{DBP}_{\mathrm{post}}$, respectively. The average of the BP values measured in step ii) and step iv) was used as the final reference value; that is, $\mathrm{SBP} = (\mathrm{SBP}_{\mathrm{pre}} + \mathrm{SBP}_{\mathrm{post}})/2$, $\mathrm{DBP} = (\mathrm{DBP}_{\mathrm{pre}} + \mathrm{DBP}_{\mathrm{post}})/2$.

The above procedure was repeated three times with a 5-minute interval to acquire three recordings on each day. In total, 12 recordings were collected for each participant on four days within one month: D (the first day), D+7, D+14, and D+21, as shown in (Fig. 2(d)). During each recording, the time interval between the smartwatch measurements and cuff BP measurements was no more than 60 seconds to ensure their consistency, in line with the recommendations of the IEEE standard for Wearable Cuffless Blood Pressure Measuring Devices (IEEE 1708) [39]. Furthermore, the synchronized ECG, PPGIR, and PPW signals were displayed in real-time on the smartwatch dial during the signal collection process, which helped the experimenter to adjust the sensor position to obtain acceptable signals. After the measurement, all recordings underwent manual double-checking to ensure the quality. A recording was considered acceptable if it was free of substantial artifacts and contained distinguishable ECG R-waves, pulse wave peaks, and valleys.

In total, 30294 recordings were collected, of which 726 were excluded due to poor signal quality. Consequently, 29568 recordings with sufficient signal quality from 3077 participants were included in subsequent analyses. Among the included participants, 1105 had a history of hypertension as indicated by previous clinical diagnosis or the use of antihypertensive drugs.

Table I summarizes the basic characteristics of the participants and compares them with the ANSI/AAMI/ISO standard. The standard requires more than 255 BP readings from at least 85 individuals to evaluate BP devices [34]. Specifically, for all

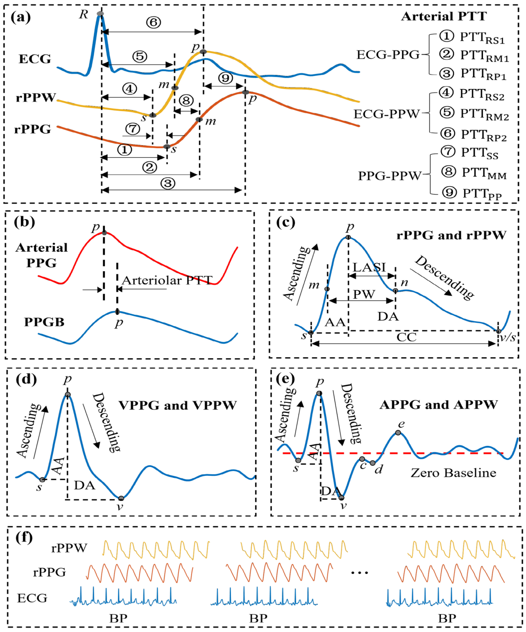

Fig. 3. Example plots of signal processing. (a) Detection of relevant characteristic points in ECG, rPPG, and rPPW signals for calculating arterial PTT. (b) Detection of arterial PPG and PPGB peaks for calculating arteriolar PTT. (c)–(e) Detection of relevant characteristic points in rPPG/rPPW, APPG/APPW, and VPPG/VPPW, respectively, to calculate handcrafted features. (f) Signal segmentation for the DL-based models.

reference SBP readings, ≥5% must be ≥160 mmHg, ≥20% must be ≥140 mmHg, and ≥5% must be ≤100 mmHg. As for all DBP readings, ≥5% must be ≥100 mmHg, ≥20% must be ≥85 mmHg, and ≥5% must be ≤60 mmHg. Table I reveals that our dataset is in high agreement with the requirements of the ANSI/AAMI/ISO standard.

The experiment was approved by the Ethics Committee of the Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences (IRB number: 210315-H0558). All participants signed informed consent before the experiment.

B. Signal Processing

The ECG and pulse wave (i.e., MWPPG and PPW) signals were preprocessed using bandpass filters with passbands of 0.5–40 Hz and 0.5–20 Hz, respectively. For each recording, characteristic points of the ECG, the raw PPG (rPPG), the first derivative of the rPPG (VPPG), the second derivative of the rPPG (APPG), the raw PPW (rPPW), the first derivative of the rPPW (VPPW), and the second derivative of the rPPW (APPW) were detected for extracting handcrafted features. First, the R (R) peak of each ECG cardiac cycle were detected (Fig. 3(a)). Second, the peak (p) points of each arterial PPG and PPGB cardiac cycle were detected (Fig. 3(b)), where the arterial PPG was extracted from PPGIR, PPGY, and PPGB by using the depth-resolved MWPPG technique proposed by Liu et al. [40]. Third, the feature points of each rPPG and rPPW cardiac cycle were detected (Fig. 3(c)), including the offset (s), maximum slope on the upward rise (m), peak (p), dicrotic notch (n), and valley (v). Fourth, the offset (s), peak (p), and valley (v) points of each VPPG and VPPW cardiac cycle were detected (Fig. 3(d)). Finally, the offset (s), peak (p), valley (v) points, and three other points (c, d, and e) of each APPG and APPW cardiac cycle were detected (Fig. 3(e)). Details for detecting characteristic points in the pulse waveform and its derivatives can be found in the previous study [41].

It should be noted that PPGIR includes arterial, arteriolar, and capillary pulses because infrared light can cross the skin and arrive at the arteries in the subcutaneous tissue [13]. Therefore, PPGIR is a mixture of PPGB (containing only capillary pulse) and PPGY (containing capillary and arteriolar pulses). To reduce redundancy among MWPPG signals, only PPGIR was used to determine rPPG during BP model construction.

C. Feature Extraction

Demographics and signal-based features employed in previous studies were used to develop the TML-based BP models, as presented in Table II. Demographics included age, gender, BMI, and history of hypertension. Signal-based features were extracted from the ECG, rPPG, APPG, VPPG, rPPW, APPW, and VPPW signals (Fig. 3). These features were classified as arterial PTT, cardiac output, or total peripheral resistance features based on their physiological mechanisms [42]. Numerous signal-based features can be derived using various signals. For example, feature *ascending time* can be obtained from the rPPG, APPG, and VPPG and from the rPPW, APPW, and VPPW. The calculation method for each signal-based feature is detailed in Table A1.

D. BP Estimation Models

Both mechanism-driven and data-driven (TML and DL) frameworks were implemented to achieve a comprehensive assessment of cuffless BP models. Specifically, three widely used PTT algorithms [as shown in (1)–(3)] were used to develop the mechanism-based BP models. The arterial PTT in (1) and (2) was set as PTT$_{RM1}$ (Table A1), which was calculated from the ECG and rPPG signals. Ridge regression, SVM, AdaBoost, and RF were chosen for the TML algorithms due to their favorable accuracy reported in the literature [16], [19], [21]. Since this study aims to evaluate the cuffless BP measurement techniques in a large-scale population, basic DL architectures commonly used in the BP literature, including VGGNet16 [22], ResNet50 [25], BiLSTM [24], and ResLSTM [25], were selected to build the DL-based models.

The PTT algorithms are suitable only for constructing calibration-based BP models, while the TML and DL algorithms can build both calibration-based and calibration-free models. A calibration-free model is a universal model trained on a population dataset that does not require modification for each individual. By contrast, a calibration-based model is a user-specific model that can be obtained by either training on an individualized dataset or by adjusting a universal model with information regarding an individual. In our study, PTT-based models were calibrated with the data of each participant on the first day and then used to predict BP on subsequent follow-up days.

To train and test the calibration-free models based on the TML and DL algorithms, we adopted a fivefold cross-validation method. All participants were randomly divided into five equal subsets. Each subset was in turn selected as the test dataset, and the remaining subsets were used for training group. The results for each fold were then combined. After building a

TABLE II
SIGNAL-BASED FEATURES AND DEMOGRAPHICS

	Features	Reference	ECG	PPG			PPW		
				rPPG	APPG	VPPG	rPPW	APPW	VPPW
Arterial PTT	ECG-PPG	[19]	✓	✓	.	.	.	.	.
	ECG-PPW	[18]	✓	.	.	.	✓	.	.
	PPG-PPW		.	✓	.	.	✓	.	.
Cardiac output (CO)	Ascending time	[43]	.	✓	✓	✓	✓	✓	✓
	Descending time		.	✓	✓	✓	✓	✓	✓
	LASI	[16]	.	✓	.	.	✓	.	.
	Pulse width	[43]	.	✓	.	.	✓	.	.
	Cardiac cycle		.	✓	.	.	✓	.	.
Total peripheral resistance (TPR)	PIR	[12]	.	✓	.	.	.	.	.
	Arteriolar PTT	[13]	.	✓	.	.	.	.	.
	Ascending slope	[44]	.	✓	✓	✓	✓	✓	✓
	Descending slope		.	✓	✓	✓	✓	✓	✓
	Ascending area		.	✓	✓	✓	✓	✓	✓
	Descending area	[36]	.	✓	✓	✓	✓	✓	✓
	AID		.	.	✓	✓	.	✓	✓
	DID		.	.	✓	✓	.	✓	✓
	Amplitudes of *p*, *v*, *c*, *d*, and *e* points	[20]	.	.	.	✓	.	.	✓
CO+TPR	Pulse K value	[17]	.	✓	.	.	✓	.	.
Demographics	Age		.	.	.	.	.	.	.
	Gender	[37]	.	.	.	.	.	.	.
	BMI		.	.	.	.	.	.	.
	History of hypertension		.	.	.	.	.	.	.

LASI, AID, and DID indicate large artery stiffness index, ascending intensity difference, and descending intensity difference, respectively.

calibration-free model, it was adjusted to create an individualized calibration-based model using the basal BP of each individual, where the basal BP is the average of BP on the first day. Let $f_{uncal}(X)$ be the calibration-free model; then the corresponding calibration-based model $f_{cal}(X)$ can be expressed as follows:

$$f_{cal}(X) = (1 - \alpha) \cdot f_{uncal}(X) + \alpha \cdot BaseBP \tag{4}$$

where X is the input features, $BaseBP$ is the base BP, and α is a balance factor. The balance factor α was determined through experimentation as the value that minimizes the estimation error of the calibration-based model.

For the PTT-based and TML-based models, the signal features listed in Table II were calculated from each cardiac cycle and then averaged within each recording. These features were then combined with the demographics listed in Table II as the input for the TML-based models. For the DL-based models, each recording (ECG, rPPG, and rPPW signals) was divided into non-overlapping 5-second segments (Fig. 3(f)), and max-min normalization was performed for each segment. The length of 5 seconds was chosen as it provides sufficient duration to capture time-domain information about cardiac activity and has shown promising results for BP estimation in previous studies [22], [27]. Additionally, the demographics listed in Table II were incorporated into the DL-based models for BP estimation using the method described in [25]. Each recording consisted of approximately 24 segments of the same reference BP value. During the evaluation phase, the BP estimates of segments from the same recording were averaged to obtain the final BP estimate for that recording.

The TML-based models were trained using the Scikit-learn library [45], and their hyperparameters were optimized using a Bayesian optimization package in Python [46]. The DL-based models were developed on a computer with eight NVIDIA Tesla K80 and using the PyTorch 1.9.1 framework. Each DL model was trained using an Adam optimizer with a learning rate of 0.001 and a batch size of 128 for 200 epochs.

The performance of the smartwatch in measuring BP was also verified by comparing the BP estimation models with the baseline models. For the calibration-based models, the baseline model was constructed by utilizing the AdaBoost algorithm with the initial BP values for calibration and demographic information as the input. For the calibration-free models, the baseline model was developed using the same algorithm with only demographic information as the input.

TABLE III
INTERNATIONAL STANDARDS AND PROTOCOLS FOR EVALUATING BP MEASUREMENTS

	ANSI/AAMI/ISO	IEEE 1708	BHS		
		MAE	CPE5	CPE10	CPE15
Grade A	-	$\leq$ 5 mmHg	60%	85%	95%
Grade B	-	5-6 mmHg	50%	75%	90%
Grade C	-	6-7 mmHg	40%	65%	85%
Grade D	-	$\geq$ 7 mmHg	Worse than Grade C		
REC†	ME $\leq$ 5 mmHg; SDE $\leq$ 8 mmHg	Grades A, B, and C	Grades A and B		

† Recommendation for clinical use.

E. Models Evaluation

The performance of the BP model was evaluated using international standards and protocols (Table III). First, the mean error (ME) and standard deviation of the error (SDE) were computed to assess the models in accordance with the ANSI/AAMI/ISO standard [34], which requires BP devices with ME and SDE values below 5 and 8 mmHg, respectively. Second, mean absolute error (MAE) was calculated to evaluate the models with regards to the IEEE 1708 standard [39], which classifies BP devices according to the MAE with various thresholds. Finally, the cumulative percentage of errors (CPE) within 5 (CPE5), 10 (CPE10), and 15 (CPE15) mmHg was calculated to to assess the models versus the British Hypertension Society (BHS) protocol [47], which grades BP devices based on their CPE at different thresholds. Statistical comparisons in the model evaluation were all two-sided and a p value less than 0.05 was considered statistically significant. Here, ME, SDE, and MAE are defined as follows:

$$ME = \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i) \tag{5}$$

TABLE IV
BP Estimation Performance of Calibration-Based and Calibration-Free Models With Various Algorithms

Algorithms		SBP		DBP	
		MAE	ME ± SDE	MAE	ME ± SDE
Calibration-based models					
Baseline		8.46	3.87 ± 11.34	5.78	2.16 ± 7.38
PTT-based	Arterial PTT [11]	9.20	4.04 ± 12.01	6.74	2.36 ± 8.53
	Arterial PTT+PIR [12]	9.26	3.92 ± 12.07	6.78	2.28 ± 8.61
	Arteriolar PTT [13]	9.26	3.90 ± 12.30	6.59	2.19 ± 8.52
TML-based	Ridge [21]	7.57	2.44 ± 9.77	5.21	1.50 ± 6.50
	SVM [17]	7.41	2.21 ± 9.62	**5.13**	**1.33** ± **6.43**
	AdaBoost [16]	**7.38**	**2.31** ± **9.57**	5.30	1.76 ± 6.52
	RF [19]	7.43	2.43 ± 9.61	5.27	1.60 ± 6.54
DL-based	VGGNet16 [22]	7.44	1.96 ± 9.67	5.22	1.27 ± 6.55
	ResNet50 [25]	7.47	2.19 ± 9.65	5.29	1.42 ± 6.59
	BiLSTM [24]	7.90	1.63 ± 10.35	5.45	1.20 ± 6.82
	ResLSTM [25]	7.55	2.11 ± 9.73	5.38	1.16 ± 6.78
Calibration-free models					
Baseline		12.19	1.75 ± 16.08	8.04	1.77 ± 10.06
TML-based	Ridge [21]	10.54	−0.39 ± 13.70	7.06	0.13 ± 9.08
	SVM [17]	10.11	−0.93 ± 13.19	**6.80**	**−0.29** ± **8.78**
	AdaBoost [16]	**9.67**	**−0.71** ± **13.04**	7.05	0.74 ± 8.93
	RF [19]	10.04	−0.41 ± 13.20	7.11	0.34 ± 9.10
DL-based	VGGNet16 [22]	10.13	−1.49 ± 13.10	7.08	−0.42 ± 9.07
	ResNet50 [25]	10.16	−0.89 ± 13.16	7.20	−0.03 ± 9.23
	BiLSTM [24]	11.68	−2.64 ± 15.27	7.69	−0.76 ± 9.89
	ResLSTM [25]	10.33	−1.13 ± 13.30	7.65	−0.68 ± 9.78

TABLE V
Performance of the Best-Performing Calibration-Based Model for Various Subpopulations

	SBP						DBP					
	MAE (mmHg)	ME (mmHg)	SDE (mmHg)	CPE5 (%)	CPE10 (%)	CPE15 (%)	MAE (mmHg)	ME (mmHg)	SDE (mmHg)	CPE5 (%)	CPE10 (%)	CPE15 (%)
All-participant	7.38	2.31	9.57	45.15	74.10	88.10	5.13	1.33	6.43	57.37	87.62	97.47
Accuracy at different BP levels												
Normotensive	**6.17**	**1.97**	**7.85**	**51.15**	**81.60**	**93.25**	**4.85**	**1.12**	**6.06**	**59.51**	**90.07**	**98.30**
Stage-1 HBP	8.10	3.42	10.07	40.61	69.48	84.87	5.05	2.06	6.17	58.43	87.35	97.45
Stage-2 HBP	8.78	2.21	11.36	38.66	66.00	82.34	5.43	1.13	6.90	54.79	84.89	96.78
Accuracy at different age levels												
(18,35]	**5.15**	**0.24**	**6.61**	**57.05**	**87.60**	**97.47**	**5.01**	**0.42**	**6.35**	**58.13**	**88.70**	**97.79**
(35,55]	7.13	1.54	9.23	45.10	75.38	89.87	5.05	1.06	6.48	58.86	88.25	97.30
(55,75]	9.21	4.42	11.12	36.47	63.21	79.86	5.28	2.21	6.34	55.65	86.34	97.37

Stage-1 HBP and Stage-2 HBP indicate stage-1 hypertension and stage-2 hypertension, respectively.

$$SDE = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - x_i - ME)^2} \tag{6}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - x_i| \tag{7}$$

where $\{x_1, x_2, \ldots x_n\}$ are the estimated BP values, $\{y_1, y_2, \ldots y_n\}$ are the reference BP values, and n is the number of BP measurements.

IV. Results

A. Results for Calibration-Based and Calibration-Free Models

Table IV presents the MAE and ME ± SDE for BP estimation errors for both calibration-based and calibration-free BP models using various algorithms. The best-performing algorithms for each model type are highlighted in bold. The balance parameter α in the models was set to 0.6 from the experimentation (Fig. A1). The best-performing calibration-based model achieved an estimation error of 2.31 ± 9.57 mmHg for SBP using the AdaBoost algorithm and 1.33 ± 6.34 mmHg for DBP using the SVM algorithm. On the other hand, the best-performing calibration-free model resulted in an estimation error of −0.71 ± 13.04 mmHg for SBP using the AdaBoost algorithm and −0.29 ± 8.78 mmHg for DBP using the SVM algorithm.

B. Subpopulation Performance Evaluations

Tables V and VI show the performance of the optimal calibration-based and calibration-free models for subpopulations with different BP categories and age levels. Following the 2017 American College of Cardiology and American Heart Association hypertension guideline [48], the BP categories were classified as normotensive (SBP less than 129 mmHg and DBP less than 79 mmHg), stage-1 hypertension (SBP at 130–139 mmHg or DBP at 80–89 mmHg), and stage-2 hypertension (SBP higher than 140 mmHg or DBP higher than 90 mmHg).

As presented in Table V, the optimal calibration-based model had the lowest estimation error for the normotensive subpopulation (1.97 ± 7.85 mmHg for SBP and 1.12 ± 6.06 mmHg for DBP) and the highest estimation error for the stage-2 hypertension subpopulation (2.21 ± 11.36 mmHg for SBP and 1.13 ± 6.90 mmHg for DBP). Similarly, the estimation error was lowest for the young subpopulation (age ≤ 35; 0.24 ± 6.61 mmHg for SBP and 0.42 ± 6.35 mmHg for DBP) and highest for the oldest subpopulation (age ≥ 55; 4.42 ± 11.12 mmHg for SBP and 2.21

TABLE VI
PERFORMANCE OF THE BEST-PERFORMING CALIBRATION-FREE MODEL FOR VARIOUS SUBPOPULATIONS

	SBP						DBP					
	MAE (mmHg)	ME (mmHg)	SDE (mmHg)	CPE5 (%)	CPE10 (%)	CPE15 (%)	MAE (mmHg)	ME (mmHg)	SDE (mmHg)	CPE5 (%)	CPE10 (%)	CPE15 (%)
All-participant	9.67	−0.71	13.04	33.33	60.19	77.79	6.80	−0.29	8.78	45.70	76.69	91.90
Accuracy at different BP levels												
Normotensive	**8.55**	**4.45**	**10.20**	**37.50**	**66.38**	**84.08**	**5.94**	**2.91**	**6.89**	**50.20**	**81.80**	**95.51**
Stage-1 HBP	9.92	−2.52	12.42	32.31	58.86	77.31	6.16	−1.75	7.56	48.72	80.31	94.65
Stage-2 HBP	12.26	−7.47	14.07	27.68	51.06	68.01	8.31	−3.71	10.07	38.18	67.75	85.43
Accuracy at different age levels												
(18,35]	**7.16**	**0.01**	**9.40**	**44.11**	**74.58**	**90.03**	**5.93**	**0.30**	**7.49**	**50.63**	**81.89**	**95.26**
(35,55]	10.71	−1.00	13.75	29.65	55.57	74.92	7.67	−1.35	9.95	41.89	71.77	88.13
(55,75]	11.40	−1.02	14.65	28.24	53.19	71.46	6.76	0.11	8.60	45.04	76.69	92.36

Stage-1 HBP and Stage-2 HBP indicate stage-1 hypertension and stage-2 hypertension, respectively.

TABLE VII
PERFORMANCE EVALUATION OF THE OPTIMAL CALIBRATION-BASED AND CALIBRATION-FREE MODELS FOR VARIOUS INTERNATIONAL STANDARDS AND PROTOCOLS UNDER SUBPOPULATIONS

	Calibration-based SBP			Calibration-free SBP			Calibration-based DBP			Calibration-free DBP		
	IEEE 1708	ANSI/ AAMI /ISO	BHS	IEEE 1708	ANSI/ AAMI /ISO	BHS	IEEE 1708	ANSI/ AAMI /ISO	BHS	IEEE 1708	ANSI/ AAMI /ISO	BHS
All-participant	Grade D	Fail	Grade C	Grade D	Fail	Grade D	Grade B	**Pass**	Grade B	Grade C	Fail	Grade C
Accuracy at different BP levels												
Normotensive	Grade C	**Pass**	Grade B	Grade D	Fail	Grade D	Grade A	**Pass**	Grade B	Grade B	**Pass**	Grade B
Stage-1 HBP	Grade D	Fail	Grade C	Grade D	Fail	Grade D	Grade B	**Pass**	Grade B	Grade C	**Pass**	Grade C
Stage-2 HBP	Grade D	Fail	Grade D	Grade D	Fail	Grade D	Grade B	**Pass**	Grade B	Grade D	Fail	Grade D
Accuracy at different age levels												
(18,35]	Grade B	**Pass**	Grade B	Grade D	Fail	Grade C	Grade B	**Pass**	Grade B	Grade B	**Pass**	Grade B
(35,55]	Grade D	Fail	Grade C	Grade D	Fail	Grade D	Grade B	**Pass**	Grade B	Grade D	Fail	Grade C
(55,75]	Grade D	Fail	Grade D	Grade D	Fail	Grade D	Grade B	**Pass**	Grade B	Grade C	Fail	Grade C

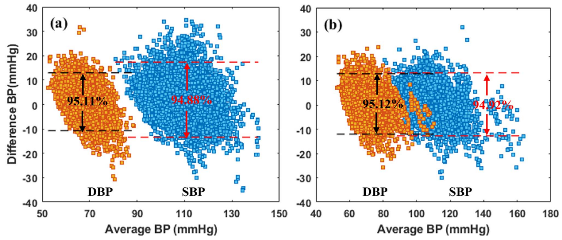

Fig. 4. Bland-Altman plots of estimated SBP from the optimal calibration-based model against the reference for (a) normotensive and (b) young subpopulations. The dotted lines in (a) and (b) represent ME ± 1.96 × SDE.

± 6.34 mmHg for DBP). The results for the optimal calibration-free model were similar (Table VI). These findings suggest that the performance of cuffless BP measurement models degrades in individuals with higher BP and age levels. Therefore, studies focusing on a small cohort of young and healthy individuals may not be generalizable to larger and more heterogeneous populations.

Fig. 4 presents Bland-Altman plots of estimated BP values from the best-performing calibration-based model compared to the reference auscultated BP values for the normotensive and young subpopulations; dashed lines indicate the 95% confidence intervals (ME ± 1.96 × SDE). The percentages of both the SBP and DBP estimates met or were close to the 95% ratio from the Bland–Altman analysis, suggesting that the smartwatch measurements were generally consistent with the reference measurements for these populations.

We compared the performance of the optimal calibration-based and calibration-free models to the ANSI/AAMI/ISO and IEEE 1708 standards, as well as the BHS protocol. Table VII shows that the performance of the models evaluated by different criteria was broadly the same in each group. The results obtained by the IEEE 1708 standard were consistent with those obtained by the ANSI/AAMI/ISO standard and the BHS protocol in terms of recommendations for clinical use. Specifically, the DBP estimates of the optimal calibration-based model were sufficiently accurate for both the overall population and different subpopulations, meeting the clinical recommendations of the ANSI/AAMI/ISO and IEEE 1708 standards and the BHS protocol. For SBP estimation, the optimal calibration-based model performed well for the normotensive and young subpopulations (satisfied the clinical recommendations of the ANSI/AAMI/ISO and IEEE 1708 standards and the BHS protocol), but not for the older individuals or individuals with hypertension. Except for DBP estimation in the normotensive and young subpopulations, the calibration-free model's performance was unsatisfactory.

C. Robustness Evaluation

Calibration is generally required for cuffless BP models to maintain the accuracy acceptable [8]. However, it is important to investigate the robustness of calibration-based models, i.e., whether the model performance degrades over the follow-up period after calibration [39]. We evaluated the absolute error of BP estimation on days 7, 14, and 21 after calibration (Fig. 5(a) and (b)). The absolute error of BP estimation increased significantly from D+7 to D+14 (mean ± SD of 7.41 ± 6.43 mmHg vs. 9.07 ± 7.24 mmHg for SBP, 5.03 ± 3.28 mmHg vs. 6.32 ± 4.18 mmHg for DBP) but remained stable from D+14 to D+21 (mean ± SD of 9.07 ± 7.24 mmHg vs. 9.31 ± 7.31 mmHg for SBP, 6.32 ± 4.18 mmHg vs. 6.24 ± 4.18 mmHg for DBP). We further analyzed the absolute change in BP relative to day D (|△BP|) on days D+7, D+14, and D+21, and found that the

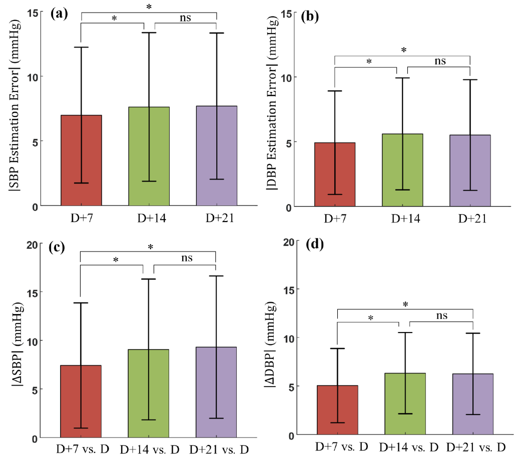

Fig. 5. Estimation errors on days D+7, D+14, and D+21 for SBP (a) and DBP (b). SBP changes (c) and DBP changes (d) on days D+7, D+14, and D+21 relative to day D. Asterisk ($*$) and "ns" indicates $p < 0.05$ and $p > 0.05$, respectively.

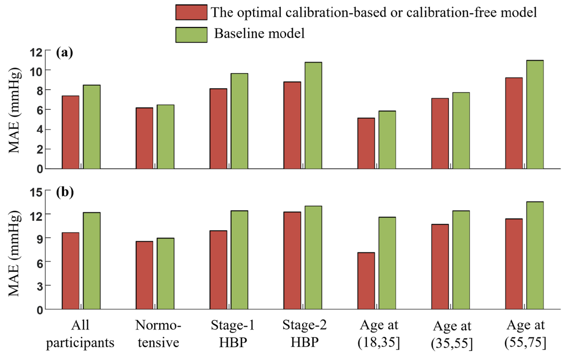

Fig. 6. Optimal calibration-based (a) and calibration-free (b) models compared to their respective baseline models for SBP estimation.

$|\triangle$ BP$|$ on D+14 was significantly greater than that on D+7, but comparable to that on D+21 (Fig. 5(c) and (d)), suggesting that model performance may fluctuate, but did not significantly decline over the 1-month period after calibration.

D. Comparison With Baseline Models

Fig. 6 shows the comparison of the optimal calibration-based and calibration-free models with the baseline models under different subpopulations, using the MAE as the evaluation metric. Since DBP has lower variation than SBP, the results of SBP estimation were presented as the example. As shown in Fig. 6, both the optimal calibration-based and calibration-free models exhibit lower errors in SBP estimation than their corresponding baseline models across all subpopulations. Notably, the optimal models demonstrate a more significant advantage over the baseline models in the hypertensive subpopulations (e.g., MAE = 8.10 mmHg *versus* MAE = 9.64 mmHg for SBP estimation at stage-1 hypertension with calibration-based strategy). These results suggest that the smartwatch can provide extra values in estimating BP, particularly for individuals with hypertension.

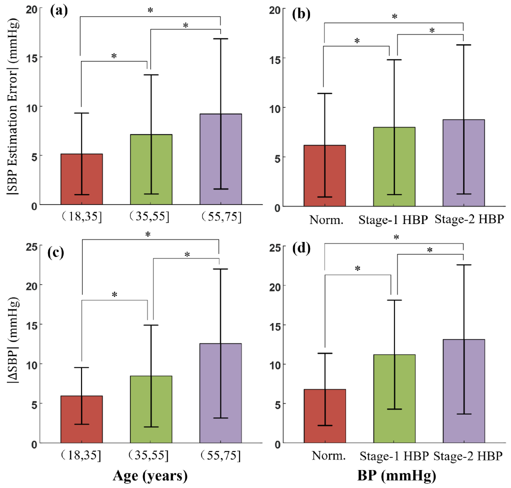

Fig. 7. SBP estimation errors for various age groups (a) and BP levels (b). SBP changes for various age groups (c) and BP levels (d). Abbreviations Norm, Stage-1 HBP, and Stage-2 HBP indicate normotensive, stage-1 hypertension, and stage-2 hypertension, respectively. Asterisk ($*$) indicates $p < 0.05$.

V. Discussion

To our knowledge, this is the largest-scale study to validate the feasibility of using smartwatches to measure BP by utilizing dual-observer auscultation BP as the reference measurement. With the calibration-based strategy, the smartwatch presented high consistency with the reference device for measuring DBP for diverse and heterogeneous participants and performed well for measuring SBP for normotensive and young participants. Smartwatch performance for both calibration-based and calibration-free BP measurements was influenced by age and BP levels. This study provided key benchmarks for future investigations of cuffless BP measurement techniques.

A. Effects of Age and BP Level on Performance

The model's performance decreased as age and BP increased (Fig. 7(a) and (b)). These findings validate a previous hypothesis [21] that because only a small cohort of young and healthy participants were enrolled in the current cuffless BP models, their performance would perform worse than initially reported when applied to heterogeneous populations. Therefore, we can conclude that if the testing dataset participants do not have the BP distribution required by the ANSI/AAMI/ISO standard, the performance results may be overly optimistic, and such studies may not have clinical utility.

By analyzing BP change (defined as the $|\triangle$BP$|$ relative to the basal BP) in different subgroups, we observed that older participants and those with hypertension tended to have greater BP variability than younger and healthier participants (Fig. 7(c) and (d)). Indeed, young and healthy participants tend to have strong reflex adaptations to stress and thus have a stable hemodynamic state. By contrast, in the older individuals or patients with hypertension, hemodynamic instability tends to occur due to decreases in arterial elasticity and the effects of antihypertensive drugs. Our results also suggest that signals that can be collected by wearables, such as ECG, PPW, and MWPPG, may not fully reflect BP changes during hemodynamic instability. Therefore,

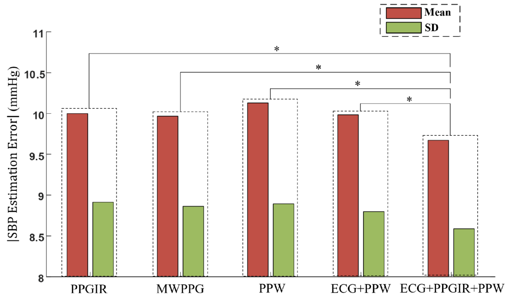

Fig. 8. Performance of AdaBoost-based calibration-free model for SBP estimation with various signal combinations. Asterisk (*) indicates $p < 0.05$.

it may be necessary to employ additional signals or principles to improve the accuracy of BP estimation for these participants.

B. Multichannel Signals for BP Estimation

Several cuffless BP measurement approaches have been reported, including those based on ECG and PPG [19], ECG and PPW [18], [31], MWPPG [13], one-channel PPG [28], and even one-channel ECG [24], [25]. However, it should be analyzed whether multichannel signals could achieve better performance to facilitate the design of wearable BP measurement devices. Therefore, we analyzed the absolute errors of SBP estimation for various signal combinations using the best-performing calibration-free model (Adaboost) as an example. The detailed results (mean ± SD) were presented in Table A2. Similar performances were observed with one-channel signal from Table A2, including PPW, PPIR, PPGR, PPGY and PPGB. However, the combination of multi-channel signals can improve the performance. Fig. 8 further illustrates the performance comparison between models based on one-channel signal and multi-channel signals, with statistical differences between pairs of groups marked with an asterisk. Compared with PPGIR- and PPW-based models, MWPPG-based model performed slightly better due to arteriolar PTT involved. However, this difference was not significant, likely due to the instability of arteriolar PTT measurements caused by changes in the location and contact force of the MWPPG sensor with the skin during long-term follow-up periods [49]. Additionally, the estimation errors were significantly reduced by fusing of ECG, PPGIR, and PPW signals, indicating that multichannel signal fusion could improve the BP estimation model performance. However, further research is necessary to investigate the trade-off between measurement performance and the cost of wearable devices.

C. Comparison of BP Modeling Algorithms

Both mechanism-based (i.e., PTT) and data-driven (i.e., TML and DL) solutions were used to develop BP models for a comprehensive assessment. The TML algorithm (with handcrafted features) demonstrated optimal performance among these three types of algorithms (Table IV). Specifically, compared with the best-performing DL-based model (VGGNet16), the best-performing TML-based model (AdaBoost) had significantly better performance for SBP estimation but comparable performance for DBP estimation. These results suggest that the extracted features with explicit physiological meaning (Table II) play a critical role in SBP estimation. However, DL algorithms have excellent potential to achieve better performance if more complex frameworks were used; our study only investigated DL algorithms with basic architectures.

PTT-based models showed decreased performance when applied to large, heterogeneous populations with a long-term follow-up period (Table IV). The primary reason for this may be that arterial PTT and arteriolar PTT reflect only one factor that induces BP changes, while the factors that influence BP changes are complex for heterogeneous populations and long-term follow-up periods [17]. Previous studies have also shown that PTT algorithms might only be suitable for short-term continuous BP tracking and require frequent calibration for long-term use [50], [51].

D. Limitations

Although this study provides a large-scale benchmark for emerging investigations on cuffless BP measurement, it has some limitations. First, although the BP distribution of the dataset met the ANSI/AAMI/ISO standard, it remains unbalanced, with a relatively small proportion of hypertensive (SBP $\geq$ 160 mmHg or DBP $\geq$ 100 mmHg) and hypotensive (DBP $\leq$ 60 mmHg) samples collected. Therefore, the application of data augmentation or balancing techniques could further improve the accuracy of BP models. Second, to provide a benchmark for emerging investigations on cuffless BP measurements, we only applied basic algorithms for BP estimation. However, more complex algorithms, such as hybrid networks that fuse TML and DL models or combine mechanism-driven and data-driven models, could potentially achieve superior performance for cuffless BP estimation. Third, the experimental procedure can be more delicate. For example: BP interventions were not included in the study due to potential ethical risks to individuals; the model's robustness was evaluated only for about a month, and longer-term follow-up is required to ensure its reliability; the recordings were collected in a controlled environment with manually controlled signal quality, the performance of the smartwatch may be affected in real-world scenarios.

VI. Conclusion

This is the largest validation study to date evaluating the performance of cuffless BP measurement techniques using smartwatches with dual-observer auscultation BP reference measurements. Our findings reveal the following: 1) Smartwatches exhibit reliable performance in estimating DBP for diverse and heterogeneous participants and SBP for normotensive and young participants with calibration, 2) The performance of smartwatches in estimating BP decreases with age and higher BP levels, 3) The availability of cuffless BP measurement without calibration is limited in routine settings, and 4) When applied to large heterogeneous populations and long follow-up periods, data-driven BP models generally outperform mechanism-driven BP models. These findings suggest that BP models trained on small cohorts of young and healthy participants may exhibit poor performance when applied to large-scale, diverse populations. The protocol and participant classification used in this study were designed in strict adherence to the ANSI/AAMI/ISO standard, and thus are likely to be generalizable in real-world settings. Overall, this study provides critical benchmarks for future investigations of emerging cuffless BP measurement techniques.

TABLE A1
DEFINITIONS OF THE SIGNAL-BASED FEATURES

Feature Type	Feature	Definitions	Feature Type	Feature	Definitions
Arterial PTT	PTT_{RS1}	Time span between ECG *R* peak and raw PPG *s* point	CO	Pulse width	Time span between points *m* and *n*
	PTT_{RM1}	Time span between ECG *R* peak and raw PPG *m* point		Cardiac cycle	Time span between points *s* and *v*
	PTT_{RP1}	Time span between ECG *R* peak and raw PPG *p* point	Total peripheral resistance (TPR)	PIR	Ratio of *p* point intensity to *s* point intensity
	PTT_{RS2}	Time span between ECG *R* peak and raw PPW *s* point		Arteriolar PTT	Time span between the peaks of arterial PPG and PPGB (Fig. 3b)
	PTT_{RM2}	Time span between ECG *R* peak and raw PPW *m* point		Ascending slope	Slope between points *s* and *p*
	PTT_{RP2}	Time span between ECG *R* peak and raw PPW *p* point		Descending slope	Slope between points *p* and *v*
	PTT_{SS}	Time span between raw PPG *s* point and raw PPW *s* point		Ascending area	Area below the curve surrounded by points *s* and *p*
	PTT_{MM}	Time span between raw PPG *m* point and raw PPW *m* point		Descending area	Area below the curve surrounded by points *p* and *v*
	PTT_{PP}	Time span between raw PPG *p* point and raw PPW *p* point		AID	Amplitude difference between points *p* and *s*
Cardiac output (CO)	Ascending time	Time span between points *s* and *p*		DID	Amplitude difference between points *p* and *v*
	Descending time	Time span between points *p* and *v*		Amplitudes of *p*, *v*, *c*, *d*, and *e* points	Amplitudes of points *p*, *v*, *c*, *d*, and *e* relative to the baseline (Fig. 3e)
	LASI	Time span between points p and n	CO+TPR	Pulse K value	Defined in [17]

APPENDIX

TABLE A2
ABSOLUTE ERRORS OF ADABOOST-BASED CALIBRATION-FREE MODEL FOR SBP ESTIMATION UNDER DIFFERENT SIGNAL COMBINATIONS

Signal combinations	Absolute Errors (mmHg)	
	mean	SD
PPGIR	9.99	8.85
PPGR	10.05	8.86
PPGY	9.98	8.89
PPGB	10.05	8.88
MWPPG	9.96	8.86
PPW	10.12	8.89
PPGIR+PPW	9.87	8.76
ECG+PPGIR	9.90	8.78
ECG+PPW	9.98	8.79
ECG+PPGIR+PPW	9.67	8.58

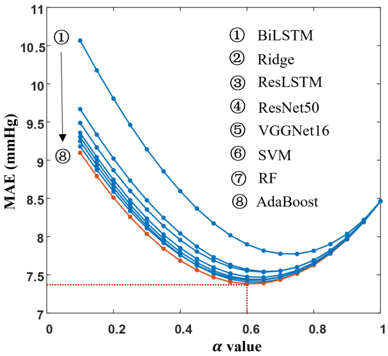

Fig. A1. MAE values of various calibration-based models at different balance parameter α (with a step size of 0.05).

REFERENCES

[1] F. D. Fuchs and P. K. Whelton, "High blood pressure and cardiovascular disease," *Hypertension*, vol. 75, no. 2, pp. 285–292, 2020.

[2] K. T. Mills, A. Stefanescu, and J. He, "The global epidemiology of hypertension," *Nature Rev. Nephrol.*, vol. 16, no. 4, pp. 223–237, 2020.

[3] World Health Organization (WHO), "Hypertension," Accessed: Nov. 20, 2022. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/hypertension

[4] E. O'brien et al., "European society of hypertension recommendations for conventional, ambulatory and home blood pressure measurement," *J. Hypertension*, vol. 21, no. 5, pp. 821–848, 2003.

[5] J. R. Banegas et al., "Relationship between clinic and ambulatory blood-pressure measurements and mortality," *New England J. Med.*, vol. 378, no. 16, pp. 1509–1520, 2018.

[6] W.-Y. Yang et al., "Association of office and ambulatory blood pressure with mortality and cardiovascular outcomes," *JAMA*, vol. 322, no. 5, pp. 409–420, 2019.

[7] R. Agarwal and R. P. Light, "The effect of measuring ambulatory blood pressure on nighttime sleep and daytime activity–implications for dipping," *Clin. J. Amer. Soc. Nephrol.*, vol. 5, no. 2, pp. 281–285, 2010.

[8] X.-R. Ding and Y.-T. Zhang, "Pulse transit time technique for cuffless unobtrusive blood pressure measurement: From theory to algorithm," *Biomed. Eng. Lett.*, vol. 9, no. 1, pp. 37–52, 2019.

[9] C. Landry, E. T. Hedge, R. L. Hughson, S. D. Peterson, and A. Arami, "Accurate blood pressure estimation during activities of daily living: A wearable cuffless solution," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 7, pp. 2510–2520, Jul. 2021.

[10] V. G. Ganti, A. M. Carek, B. N. Nevius, J. A. Heller, M. Etemadi, and O. T. Inan, "Wearable cuff-less blood pressure estimation at home via pulse transit time," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 6, pp. 1926–1937, Jun. 2021.

[11] C. C. Y. Poon and Y.-T Zhang, "Cuff-less and noninvasive measurements of arterial blood pressure by pulse transit time," in *Proc. IEEE 27th Annu. Conf. Eng. Med. Biol.*, 2005, pp. 5877–5880.

[12] X.-R. Ding, Y.-T. Zhang, J. Liu, W.-X. Dai, and H. K. Tsang, "Continuous cuffless blood pressure estimation using pulse transit time and photoplethysmogram intensity ratio," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 5, pp. 964–972, May 2016.

[13] J. Liu, B. P. Yan, Y.-T. Zhang, X.-R. Ding, P. Su, and N. Zhao, "Multi-wavelength photoplethysmography enabling continuous blood pressure measurement with compact wearable electronics," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 6, pp. 1514–1525, Jun. 2019.

[14] F. Miao, B. Zhou, Z.-D Liu, B. Wen, Y. Li, and M. Tang, "Using noninvasive adjusted pulse transit time for tracking beat-to-beat systolic blood pressure during ventricular arrhythmia," *Hypertension Res.*, vol. 45, no. 3, pp. 424–435, 2022.

[15] Y. Chen, S. Shi, Y.-K. Liu, S.-L. Huang, and T. Ma, "Cuffless blood-pressure estimation method using a heart-rate variability-derived parameter," *Physiol. Meas.*, vol. 39, no. 9, 2018, Art. no. 095002.

[16] 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.

[17] F. Miao et al., "A novel continuous blood pressure estimation approach based on data mining techniques," *IEEE J. Biomed. Health Inform.*, vol. 21, no. 6, pp. 1730–1740, Nov. 2017.

[18] F. Miao, Z.-D. Liu, J.-K. Liu, B. Wen, Q.-Y. He, and Y. Li, "Multi-sensor fusion approach for cuff-less blood pressure measurement," *IEEE J. Biomed. Health Inform.*, vol. 24, no. 1, pp. 79–91, Jan. 2020.

[19] S. Yang, J. Sohn, S. Lee, J. Lee, and H. C. Kim, "Estimation and validation of arterial blood pressure using photoplethysmogram morphology features in conjunction with pulse arrival time in large open databases," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 4, pp. 1018–1030, Apr. 2021.

[20] S. Haddad, A. Boukhayma, and A. Caizzone, "Continuous PPG-based blood pressure monitoring using multi-linear regression," *IEEE J. Biomed. Health Inform.*, vol. 26, no. 5, pp. 2096–2105, May 2022.

[21] R. Mieloszyk et al., "A comparison of wearable tonometry, photoplethysmography, and electrocardiography for cuffless measurement of blood pressure in an ambulatory setting," *IEEE J. Biomed. Health Inform.*, vol. 26, no. 7, pp. 2864–2875, Jul. 2022.

[22] Z.-D Liu, F. Miao, R.-X Wang, J.-K Liu, B. Wen, and Y. Li, "Cuff-less blood pressure measurement based on deep convolutional neural network," in *Proc. IEEE 41st Annu. Int. Conf. Eng. Med. Biol. Soc.*, 2019, pp. 3775–3778.

[23] 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 Inform.*, 2018, pp. 323–328.

[24] X.-M Fan, H. Wang, F. Xu, Y. Zhao, and K.-L. Tsui, "Homecare-oriented intelligent long-term monitoring of blood pressure using electrocardiogram signals," *IEEE Trans. Ind. Informat.*, vol. 16, no. 11, pp. 7150–7158, Nov. 2020.

[25] F. Miao et al., “Continuous blood pressure measurement from one-channel electrocardiogram signal using deep-learning techniques,” *Artif. Intell. Med.*, vol. 108, 2020, Art. no. 101919.

[26] 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 Inform.*, vol. 26, no. 5, pp. 2075–2085, May 2022.

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

[28] D.-K. Kim, Y.-T. Kim, H. Kim, and D.-J. Kim, “DeepCNAP: A deep learning approach for continuous noninvasive arterial blood pressure monitoring using photoplethysmography,” *IEEE J. Biomed. Health Inform.*, vol. 26, no. 8, pp. 3697–3707, Aug. 2022.

[29] S. Baker, W. Xiang, and I. J.K.-B. S. Atkinson, “A computationally efficient CNN-LSTM neural network for estimation of blood pressure from features of electrocardiogram and photoplethysmogram waveforms,” *Knowl.-Based Syst.*, vol. 250, 2022, Art. no. 109151.

[30] X.-R. Ding et al., “Continuous blood pressure measurement from invasive to unobtrusive: Celebration of 200th birth anniversary of Carl Ludwig,” *IEEE J. Biomed. Health Inform.*, vol. 20, no. 6, pp. 1455–1465, Nov. 2016.

[31] Z.-D. Liu, J.-K. Liu, B. Wen, Q.-Y. He, Y. Li, and F. Miao, “Cuffless blood pressure estimation using pressure pulse wave signals,” *Sensors*, vol. 18, no. 12, 2018, Art. no. 4227.

[32] R. Mukkamala et al., “Evaluation of the accuracy of cuffless blood pressure measurement devices: Challenges and proposals,” *Hypertension*, vol. 78, no. 5, pp. 1161–1167, 2021.

[33] R. Mukkamala, S. G. Shroff, C. Landry, K. G. Kyriakoulis, A. P. Avolio, and G. S. Stergiou, “The microsoft research aurora project: Important findings on cuffless blood pressure measurement,” *Hypertension*, vol. 80, no. 3, pp. 534–540, 2023.

[34] *ANSI/AAMI/ISO 81060-2:2019*, “Non-invasive sphygmomanometers - Part 2: Clinical investigation of intermittent automated measurement type,” Accessed: May 15, 2023. [Online]. Available: https://webstore.ansi.org/Standards/AAMI/ANSIAAMIISO810602019

[35] Y. Ma et al., “Relation between blood pressure and pulse wave velocity for human arteries,” *Proc. Nat. Acad. Sci.*, vol. 115, no. 44, pp. 11144–11149, 2018.

[36] W.-H. Lin, H. Wang, O. W. Samuel, G. Liu, Z. Huang, and G. Li, “New photoplethysmogram indicators for improving cuffless and continuous blood pressure estimation accuracy,” *Physiol. Meas.*, vol. 39, no. 2, 2018, Art. no. 025005.

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

[38] R. Padwal et al., “Optimizing observer performance of clinic blood pressure measurement: A position statement from the Lancet Commission on Hypertension Group,” *J. Hypertension*, vol. 37, no. 9, pp. 1737–1745, 2019.

[39] I. S. Association, *IEEE Standard for Wearable Cuffless Blood Pressure Measuring Devices*, IEEE Standard 708-2014, 2014.

[40] J. Liu, B. P.-Y. Yan, W.-X. Dai, X.-R. Ding, Y.-T. Zhang, and N. Zhao, “Multi-wavelength photoplethysmography method for skin arterial pulse extraction,” *Biomed. Opt. Exp.*, vol. 7, no. 10, pp. 4313–4326, 2016.

[41] M. Z. Suboh, R. Jaafar, N. A. Nayan, N. H. Harun, and M. S. F. Mohamad, “Analysis on four derivative waveforms of photoplethysmogram (PPG) for fiducial points detection,” *Front. Public Health*, vol. 10, 2022, Art. no. 920946.

[42] W.-H. Lin, X. Li, Y. Li, G. Li, and F. J. P. M. Chen, “Investigating the physiological mechanisms of the photoplethysmogram features for blood pressure estimation,” *Physiol. Meas.*, vol. 41, no. 4, 2020, Art. no. 044003.

[43] Z.-D Liu, B. Zhou, Y. Li, M. Tang, and F. Miao, “Continuous blood pressure estimation from electrocardiogram and photoplethysmogram during arrhythmias,” *Front. Physiol.*, vol. 11, 2020, Art. no. 575407.

[44] Y. S. Putyatina, “Measurement of arterial blood pressure by processing pulse wave data,” in *Proc. IEEE 3rd Annu. Siberian Russian Workshop Electron Devices Mater.*, 2002, pp. 77–78.

[45] F. Pedregosa et al., “Scikit-learn: Machine learning in python,” *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[46] F. Nogueira, “Bayesian optimization: Open source constrained global optimization tool for python,” Accessed: Sep. 25, 2022. [Online]. Available: https://github.com/fmfn/BayesianOptimization

[47] E. O’Brien et al., “The British Hypertension Society protocol for the evaluation of blood pressure measuring devices,” *J. Hypertension*, vol. 11, no. Suppl 2, pp. S43–S62, 1993.

[48] 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,” *J. Amer. College Cardiol.*, vol. 71, no. 19, pp. e127–e248, 2018.

[49] X. F. Teng and Y. T. Zhang, “Theoretical study on the effect of sensor contact force on pulse transit time,” *IEEE Trans. Biomed. Eng.*, vol. 54, no. 8, pp. 1490–1498, Aug. 2007.

[50] B. McCarthy, C. Vaughan, B. O’flynn, A. Mathewson, and C. Ó. Mathúna, “An examination of calibration intervals required for accurately tracking blood pressure using pulse transit time algorithms,” *J. Hum. Hypertension*, vol. 27, no. 12, pp. 744–750, 2013.

[51] R. Mukkamala and J.-O. Hahn, “Toward ubiquitous blood pressure monitoring via pulse transit time: Predictions on maximum calibration period and acceptable error limits,” *IEEE Trans. Biomed. Eng.*, vol. 65, no. 6, pp. 1410–1420, Jun. 2018.