diagnostics

MDPI

Article

Lightweight Calibration-Based Blood Pressure Tracking from Photoplethysmography Features for Smart Devices: Accuracy, Model Size and Recalibration Interval

Soojeong Lee [1], Mugahed A. Al-antari [2] and Gangseong Lee [3,*]

[1] Department of Computer Engineering, Sejong University, 209 Neungdong-ro, Gwangjin-gu, Seoul 05006, Republic of Korea; leesoo86@sejong.ac.kr
[2] Department of Artificial Intelligence and Data Science, College of AI Convergence, Daeyang AI Center, Sejong University, 209 Neungdong-ro, Gwangjin-gu, Seoul 05006, Republic of Korea; en.mualshz@sejong.ac.kr
[3] Ingenium College, Kwangwoon University, 20 Kwangwoon-ro, Nowon-gu, Seoul 01897, Republic of Korea
* Correspondence: gslee0115@gmail.com

Abstract

Background/Objectives: Smart devices estimate blood pressure (BP) from photoplethysmography (PPG), but models trained on other people generalize poorly, and device constraints limit model size. This study compared calibration strategies and model sizes for PPG-based BP estimation. **Methods:** We used 65 hand-crafted PPG features with MIMIC-BP (1295 training and 229 subject-disjoint test subjects, ten repeats) and 17 external VitalDB surgical cases. We compared calibration-free regressors, calibration with one to five cuffs, recalibration policies and calibration models of decreasing size. **Results:** Without calibration, six regressors reached a systolic BP (SBP) mean absolute error (MAE) of 13.0–13.3 mmHg on unseen subjects, close to a constant predictor (14.2 mmHg). A delta model that predicts the change from the cuff reading reduced the SBP MAE to 11.8 mmHg with one cuff and 10.8 mmHg with five, but followed only 30–50% of large changes. With five calibration readings, the model achieved a pooled area under the curve (AUC) of 0.94 for new-onset SBP ≥160 mmHg on MIMIC-BP, but sensitivity was zero at an estimated-SBP threshold of 160 mmHg. Accuracy was governed by the recalibration interval: recalibration nominally every 15 min gave 10.3 mmHg on VitalDB, and event-triggered cuffs were not better than periodic ones. A 47 KB tree model lost 0.2 mmHg against a 1.5 MB reference. **Conclusions:** For these data, calibration was required, the recalibration interval mattered more than model complexity, and a compact tree model was sufficient; on-device use requires further validation. Accuracy remained poor at blood-pressure extremes, and external validation in only 17 VitalDB cases remains preliminary.

Keywords: cuffless blood pressure; photoplethysmography; wearable; calibration; lightweight model; recalibration interval; trend tracking

Check for updates

Academic Editor: Kaan Orhan

Received: 20 October 2026
Revised: 5 November 2026
Accepted: 5 November 2026
Published: 10 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Hypertension affects more than a billion adults and is the leading modifiable risk factor for cardiovascular disease [1]; current guidelines emphasize out-of-office and repeated measurements for diagnosis and management [2,3]. Cuff-based devices provide only intermittent readings, which has motivated the cuffless estimation of blood pressure (BP) from signals that wearable and mobile devices can record continuously—above all, the photoplethysmogram (PPG) [4,5]. PPG-based approaches range from pulse-transit-time

models [6] to machine learning on hand-crafted pulse-wave features [7–9] and deep networks on the raw waveform [10,11]; many recent studies report errors well within clinical tolerances [12–17].

Whether these accuracies transfer to a person the model has never seen is less clear. When segments of the same subject appear in both training and test data, a model can recognize the subject rather than estimate the pressure, and reported accuracy drops markedly once such subject information leakage is prevented [18]. Pulse transit and pulse-shape features change with pressure within a person. Still, their relation to absolute pressure differs between people, which is why transit-time methods require individual calibration [6,19,20]. Accordingly, many cuffless devices need an initial and repeated (e.g. daily, weekly, or monthly) cuff calibration to report pressure in mmHg, and their accuracy depends on this calibration and its renewal [21]; dedicated standards now govern the evaluation of cuffless and continuous devices [22,23].

For a smart device, three further constraints matter. First, every cuff reading burdens the user, so the number of cuffs and the recalibration interval are design parameters that should be chosen based on evidence. Second, the estimator runs on a microcontroller or a phone with limited memory and energy, so the model must be small. Third, a device meets users and conditions not in its training data, so the estimator must remain stable under changes in population and acquisition system. Multimodal physiological sensing addresses such variability by combining several signals and weighting them by their estimated reliability, as recently shown for the fusion of acoustic signals and electrical impedance tomography in respiratory assessment [24]; a single-PPG device cannot fall back on a second modality, so signal quality has to be handled within the PPG pipeline itself. Despite their practical importance, these questions are rarely addressed together: calibrated methods are usually evaluated with a single calibration and without external data, and model size is seldom reported.

In this work, we address these questions for a PPG-only pipeline based on hand-crafted features, using a large intensive-care database with a strictly subject-disjoint design and an external surgical database. Our contributions are the following:

1. We quantify the calibration-free ceiling on unseen subjects for six regressors of very different complexity and show that it is close to a constant predictor, especially on external data.
2. We compare four calibration methods with one to five cuffs and show that a delta model, which predicts the change from the cuff, is consistently the most accurate, while characterizing its limits in following large pressure changes.
3. We derive recalibration-interval guidance from the error as a function of the cuff rate, show with a pre-specified comparison that event-triggered cuffs do not improve on periodic recalibration, and explain this by the slow, drift-like structure of the residual error.
4. We show that the delta model can be reduced about 30-fold in memory, to about 47 KB and 1200 comparisons per estimate, with a loss of about 0.2 mmHg. In contrast, random forests are impractically large, and linear or small neural models fail when the database changes.

2. Materials and Methods

Figure 1 summarizes the data, features, estimation methods and three design questions examined.

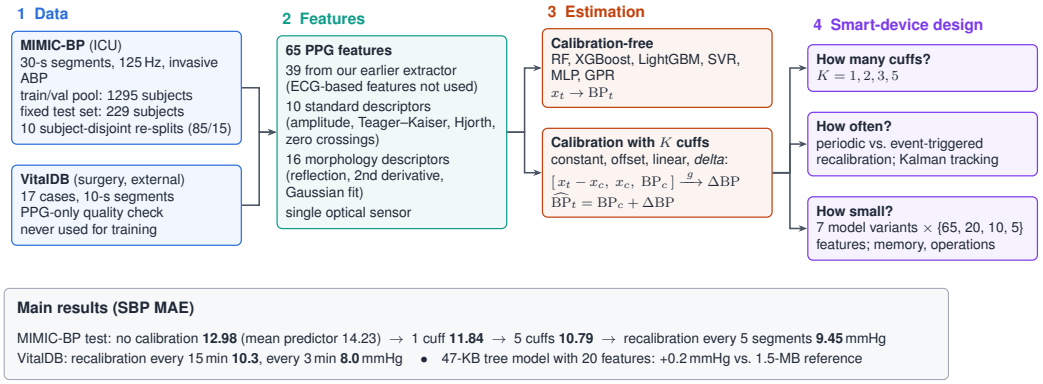


Figure 1. Study overview. PPG features from an intensive-care database (MIMIC-BP; subject-disjoint training, validation and test sets) and an external surgical database (VitalDB) were used to estimate blood pressure without calibration and with K cuff readings. The delta model g estimates the change since a cuff reading BP_c from the current features x_t and the features x_c at the cuff. The calibrated estimate was then examined with respect to the number of cuffs, the recalibration interval and the model size. Errors are SBP MAE.

2.1. Datasets and Splits

Two public databases with invasive arterial blood pressure (ABP) as reference were used. MIMIC-BP [25], derived from MIMIC-III [26], provides 30 s segments (125 Hz) of simultaneously recorded photoplethysmogram (PPG) and ABP from intensive-care patients, with 30 non-overlapping segments per subject in temporal order and a predefined subject-disjoint division into training, validation and test subjects. The test set (229 subjects) was kept fixed and untouched until all settings had been frozen. To quantify the variability due to the choice of training subjects, the training and validation subjects were pooled and re-split ten times into 85 % training and 15 % validation subjects (repeat r uses random seed r), and we selected every model and setting on the validation subjects of the same repeat.

VitalDB [27] served as an external test set from a different population and acquisition system (surgical patients). Seventeen cases with continuous PPG (SNUADC PLETH) and radial arterial pressure were resampled to 125 Hz and cut into consecutive, non-overlapping 10 s segments. Segments with missing samples, with implausible reference values (SBP outside 50–230 mmHg, DBP outside 30–140 mmHg, or pulse pressure below 15 mmHg) or with a flat PPG were discarded, and each PPG segment was standardized to zero mean and unit variance. A PPG-only quality check, blind to the blood pressure, then removed segments whose dominant frequency (0.5–8 Hz band) lay outside 0.67–3 Hz (40–180 beats/min), whose spectral peak held less than 15 % of the in-band power, whose standardized amplitude exceeded 6, or in which constant runs of at least 0.5 s covered 10 % or more of the segment; 97.0 % of the segments passed. No VitalDB data were used for training or model selection, except where stated for leave-one-case-out tuning of secondary analyses (Section 2.5). In VitalDB, the reference SBP and DBP of a segment were the means of the beat-wise systolic and diastolic values reported by the patient monitor (Solar 8000) over the segment; in MIMIC-BP, they were the medians of the beat-wise systolic and diastolic values of the ABP over each 30 s segment [25]. In both databases, we analyzed only segments in which the feature extractor processed at least one beat (usable segments). MIMIC-BP provides 30 segments for each subject (38,850 segments in the training and validation pool and 6870 in the test set); 2183 (5.6 %) and 356 (5.2 %) of them, respectively, yielded no valid beat and were excluded, so that subjects contributed 28.3 segments on average (median 30; Table 1). In VitalDB, 2597 of 27,375 segments (9.5 %) yielded no valid beat. Because the original extractor was written for multimodal recordings, it also required an ECG channel

with at least four detected R peaks. All segments of both databases contained an ECG channel, and an independent R-peak detector found at least four R peaks in all excluded MIMIC-BP test and validation segments and in 98 % of the excluded VitalDB segments; the exclusions were therefore caused almost entirely by the PPG beat delineation and feature checks, not by the ECG. The mean reference SBP of excluded and retained segments was similar in MIMIC-BP (111.9 vs. 113.1 mmHg in the test set) and lower in VitalDB (106.6 vs. 115.1 mmHg). The ECG was used only for this selection; no feature was computed from it, and a device would apply the PPG checks alone (Section 4). Table 1 summarizes the data.

Table 1. Dataset characteristics (usable segments: at least one detected beat; VitalDB after input quality control). Groups by reference SBP; numbers are segments (subjects); pressures in mmHg.

	MIMIC-BP Train + Val	MIMIC-BP Test	VitalDB
Segments/subjects	36,667/1295	6514/229	24,257/17
Segments per subject, median (mean) [b]	30 (28.3)	30 (28.4)	1276 (1458)
SBP, mean ± SD (range)	112.7 ± 18.0 (64–185)	113.1 ± 17.7 (67–186)	115.0 ± 18.6 (55–230)
DBP, mean ± SD (range)	58.5 ± 12.3 (31–116)	59.4 ± 11.8 (31–109)	59.5 ± 10.7 (30–136)
Within-subject SBP range/SD [b]	43.9/11.4	44.1/10.9	105.1/15.1
SBP <90	3267 (650)	551 (107)	1128 (16)
SBP 90–140	30,674 (1289)	5491 (228)	20,858 (17)
SBP 140–160	2429 (507)	426 (88)	1686 (16)
SBP 160–180	280 (86)	43 (14)	430 (11)
SBP ≥180	17 (7)	3 (2)	155 (7)
New-onset crisis subjects [a]	70	12	11

[a] First usable segment SBP <160 mmHg and a later segment ≥160 mmHg. [b] Over subjects; MIMIC-BP provides 30 segments per subject before the beat check.

2.2. PPG Features for On-Device Use

Features were computed from the PPG alone, so that the pipeline requires a single optical sensor. Features were computed from the first 10 s (1250 samples) of every segment in both databases, so that the feature window had the same length; in MIMIC-BP, the reference values refer to the whole 30 s segment (Section 2.1), which adds a small, unquantified error to the MIMIC-BP results, whereas in VitalDB the feature window and the reference cover the same 10 s. Beats were delineated on the PPG (PPG-anchored windows), and the beat-level feature values were summarized by their median over the beats of a segment. Three groups of features were combined (Table A1).

Previously published features (39). From the features of our earlier multimodal extractor [28], we used the 39 that are derived from the PPG alone: temporal features (systolic and diastolic times, crest time and the systolic and diastolic widths at 10–75 % of the amplitude), amplitude and intensity features (pulse intensity ratio, k value, maximum and minimum amplitude), first- and second-derivative (velocity and acceleration PPG) features [8], pulse-wave indices (augmentation index, large-artery stiffness index, inflection-point area ratio), area features, the maximum of the multitaper power spectral density, and the skewness, kurtosis and entropy of the waveform. Its ECG-based features (heart rate and three pulse arrival times) were not used, because they need a second sensor and because the ECG and PPG channels of the databases were not reliably time-aligned.

Standard signal descriptors (10). We added ten simple descriptors of the PPG: the pulse amplitude (systolic peak minus diastolic foot), the amplitudes at the systolic peak and at the diastolic foot, the mean amplitude over the beat, the maximum of the periodogram, the mean Teager–Kaiser energy [29], the three Hjorth parameters (activity, mobility and complexity) [30] and the zero-crossing rate of the mean-removed signal.

Morphology descriptors (16). To describe the pulse contour more completely, we implemented 16 further PPG descriptors following their published definitions: the reflection index and the systolic-to-diastolic peak time [31], the relative dicrotic-notch height, the pulse width and up-stroke time at 50 % amplitude, the second-derivative ratios b/a, c/a,

d/a, e/a and the aging index [8,32], the parameters of a two-Gaussian decomposition of the pulse (amplitude ratio, center distance, widths and goodness of fit) and the beat-to-beat SD of the reflection index. The resulting 65 features ($39 + 10 + 16$) were median-imputed with training statistics where a model could not handle missing values.

2.3. Calibration-Free Reference Models

To establish how far PPG features carry absolute blood pressure to unseen users, six regressors were trained to map the 65 features of a segment directly to SBP and DBP: random forest (RF; 500 trees, minimum leaf size 5, one third of the features per split), XGBoost [33] (learning rate 0.03, depth 6, up to 3000 trees with early stopping on the validation subjects), LightGBM, support vector regression with a radial kernel (fitted on a random subset of 10,000 segments), a multilayer perceptron (two hidden layers of 256 units, early stopping) and a sparse variational Gaussian process with automatic relevance determination (1000 inducing points). A model predicting the training mean served as the trivial baseline.

2.4. Calibration Models

Calibration uses K cuff readings BP_c of the same user at segments $c \in C$, $|C| = K$, $K \in \{1, 2, 3, 5\}$, taken as the first K usable segments of each subject. Cuff readings were simulated by the invasive reference values of these segments, so that the error in a real cuff measurement is not included in the main analysis. To quantify the effect of cuff measurement error, the analysis was repeated with every simulated cuff reading (each calibration and each recalibration cuff) replaced by the invasive value plus an independent Gaussian error with SD 5 or 10 mmHg (zero mean, and, in a sensitivity analysis, with an additional constant bias of +5 mmHg), applied to SBP and DBP separately; the models were not changed, and errors were always computed against the invasive reference. For this analysis, the models of the ten repeats were re-trained with identical settings on a CPU (five error draws per repeat); without cuff error, they differed from the original results by at most 0.08 mmHg. To compare all K on identical data, every method was evaluated on the segments after the fifth usable segment. Because the first K segments end closer to the evaluated segments when K is larger, the number of cuffs and the recency of the last cuff are confounded in this design. To separate them, the delta model and constant calibration were also evaluated with the calibration sets $\{5\}$, $\{4, 5\}$, $\{3, 4, 5\}$ and $\{1, \ldots, 5\}$ (positions among the usable segments of each subject), which all end at the fifth segment, on the same evaluation segments (models re-trained with identical settings, ten repeats). Four methods were compared:

- *constant*: the mean cuff value is carried forward.
- *offset*: $\widehat{\mathrm{BP}}_t = f(x_t) + \overline{\mathrm{BP}_c - f(x_c)}$, where f is the calibration-free XGBoost model.
- *linear*: $\widehat{\mathrm{BP}}_t = a + b f(x_t)$ with the slope b shrunk towards 1 by a ridge penalty ($\lambda = 1000$), fitted on the $K \geq 2$ cuffs.
- *delta*: a model g maps the pair $[x_t - x_c,\ x_c,\ \mathrm{BP}_c]$ to the change $\mathrm{BP}_t - \mathrm{BP}_c$, and $\widehat{\mathrm{BP}}_t = \frac{1}{K} \sum_{c \in C} (\mathrm{BP}_c + g(x_t, x_c, \mathrm{BP}_c))$.

The delta model was trained on pairs of segments of the same training subject, with two randomly chosen reference segments per target segment, so that it applies to any calibration point and any time since calibration. The approach is similar in spirit to the estimation of blood pressure changes from differences in PPG features relative to a baseline measurement [13]; here, the reference features and pressure are also provided to the model, so that it can learn how the change depends on the starting point. Its default implementation is XGBoost with the settings above (early stopping on validation pairs).

2.5. Recalibration and Tracking

Periodic recalibration replaced the reference by a new cuff every m segments (MIMIC-BP: m = 2–15 segments; VitalDB: 3–60 min), using the most recent reference for all segments until the next cuff. In VitalDB, the interval was counted in retained 10 s segments (e.g., 90 segments for 15 min), so it equals or exceeds the nominal clock time where segments had been excluded. *Event-triggered* policies requested a cuff when the predicted SBP change since the last cuff exceeded a threshold $\tau \in \{5, \ldots, 30\}$ mmHg (with a refractory period), optionally combined with a periodic fallback, and an optional subject-specific gain on the predicted change was estimated from the user's previous cuffs with shrinkage towards 1. These policies were simulated online: each segment was scored with the estimate available before the policy decided on a cuff, so that all policies were scored on the same segments. They were compared with periodic recalibration at the same or a lower cuff rate (reference: every 5 segments in MIMIC-BP, every 15 min in VitalDB), with the configuration selected on the MIMIC-BP validation subjects or, for VitalDB, by leave-one-case-out; an adaptive policy was to be declared better only if it reduced SBP MAE by at least 0.5 mmHg with a 95 % confidence interval excluding zero (rule fixed before the analysis). Finally, causal Kalman filters (states: BP, or BP and bias) and an unscented Kalman filter (BP, bias and gain) fused the delta estimates with the cuffs; measurement noise was estimated on validation residuals and process noise tuned on validation (VitalDB: leave-one-case-out). The autocorrelation of the delta-model residuals was computed to relate the error dynamics to the recalibration interval.

2.6. Lightweight-Model Setup

To assess suitability for a wearable device, the delta model was re-implemented with models of decreasing size: the reference XGBoost; a small XGBoost (depth 3, up to 200 trees, learning rate 0.1); a small LightGBM (15 leaves, up to 200 trees); RF with 300 trees and a small RF (50 trees, depth 8, minimum leaf 20); ridge regression; and a one-hidden-layer MLP with 32 units. Each was trained with all 65 features and with the 20, 10 and 5 most important features, ranked by the gain importance of the reference model on the training pairs of the same repeat. The cost of the complete pipeline (pre-processing, feature extraction and model) was assessed by timing a Python 3.11.18 re-implementation of the feature extraction per 10 s segment and by re-training the small model without the five Gaussian-decomposition features, whose iterative curve fitting is the most expensive step. Model complexity was quantified by the number of parameters (tree nodes or weights), an estimate of the on-device memory (8 bytes per tree node, 4 bytes per weight), the number of operations per estimate (comparisons along one path per tree, or multiply–adds) and the single-thread latency of one estimate measured in Python on a desktop CPU, which serves only as a relative reference and not as an estimate of embedded latency. Before the analysis, we fixed the rule to select the smallest model whose validation SBP MAE with one cuff was within 0.3 mmHg of the reference, and to call it non-inferior on the MIMIC-BP test set if the upper 95 % confidence bound of the MAE difference was below 0.5 mmHg.

2.7. Evaluation and Statistics

Estimation errors $e = \widehat{\mathrm{BP}} - \mathrm{BP}$ were summarized by the mean absolute error (MAE), the mean error and its SD (ME ± SD), and the British Hypertension Society cumulative percentages within 5, 10 and 15 mmHg [34]; because the mean errors were close to zero, the root-mean-square error (RMSE) equals the SD and is given only where it aids comparison with other studies. The AAMI/ESH/ISO criteria [35,36] are reported only as a reference, because the protocol of a retrospective analysis does not meet the requirements of the standard; the standards for cuffless and continuous devices [22,23] likewise require prospective

protocols. The ability to follow a rise into the hypertensive range was assessed as the area under the ROC curve (AUC) for detecting segments with SBP ≥160 mmHg among segments of subjects whose calibration SBP was below 160 mmHg (new-onset), and the number of such subjects is reported. The tracking slope was obtained by regressing the estimated on the true change since the cuff. Point estimates are means over the ten repeats. Throughout the revised manuscript, all 95 % confidence intervals (CIs) were obtained by a subject-level (VitalDB: case-level) cluster bootstrap with 2000 resamples, in which the repeats were kept inside each resample, so they describe the uncertainty due to the sample of test subjects; the variability due to the choice of training subjects is reported separately as the SD over the ten repeats where relevant. Differences between methods were bootstrapped on the same resamples (paired). For the new-onset analysis, we additionally report the number of positive segments and subjects, and the sensitivity and specificity at an alert threshold of 160 mmHg on the estimate. False alarms per hour were computed from two thresholds, which are distinct: a segment with reference SBP <160 mmHg (the new-onset threshold) and estimated SBP ≥140 mmHg (the alarm threshold) is a false positive (FP); one with reference SBP <160 mmHg and estimated SBP <140 mmHg is a true negative (TN). With segment duration d in seconds, the false-positive rate per hour is $3600\,\mathrm{FP}/((\mathrm{FP}+\mathrm{TN})\,d)$ (30 s for MIMIC-BP, 10 s for VitalDB), so the denominator is the cumulative monitored time of segments whose reference SBP had not reached 160 mmHg. This is a rate of segments misclassified as positive, not a count of independent alarm events; because the estimation residuals are autocorrelated (Section 3.7), such misclassified segments are likely to cluster in time rather than occur independently, so the reported rate describes an average burden over the monitored time at risk rather than how alarms would be spaced during continuous use. The analysis was implemented in Python (XGBoost, LightGBM, scikit-learn, PyTorch 2.4.1) and MATLAB R2025b. The analysis code, the model settings and a description of the pre-processing are provided as Supplementary Material S1 for review and will be released publicly with the article (Data Availability Statement).

3. Results

3.1. Calibration-Free Estimation Reaches a Ceiling

Without calibration, all six regressors reached almost the same accuracy on the unseen MIMIC-BP test subjects (Table 2): SBP MAE ranged from 12.98 (LightGBM) to 13.25 mmHg (SVR) (SD over the ten training repeats 0.02–0.06 mmHg), only about 1.2 mmHg below a model that always predicts the training mean (14.23 mmHg). Differences between the tree ensembles and the Gaussian process were at most 0.13 mmHg, and the BHS percentages within 5, 10 and 15 mmHg (24, 46 and 65 %) were identical across models and hardly better than the mean predictor (21, 41 and 59 %). On the external VitalDB cases, the calibration-free models did not improve on the mean predictor at all (SBP MAE 14.09–14.87 vs. 14.06 mmHg). The error depended strongly on the reference pressure (Figure 2b): estimates regressed towards the population mean, with a mean error of about −42 mmHg at SBP 160–170 mmHg. The sparse high-pressure range contributes to this (Figure 2a, Table 1), but the model choice did not change it, indicating that the PPG features carry little user-independent information on absolute pressure.

3.2. Calibration with One to Five Cuffs

A single cuff reading combined with the delta model reduced the SBP MAE on the MIMIC-BP test subjects from 12.99 to 11.84 mmHg [95 % CI 11.21, 12.48] and the DBP MAE from 8.19 to 6.84 mmHg. Additional cuffs improved the estimate further (three cuffs: 11.11 mmHg; five cuffs: 10.79 [10.29, 11.30] mmHg SBP, 5.98 mmHg DBP). The delta model was better than offset calibration for every number of cuffs (SBP difference −2.51

[−3.32, −1.71] mmHg with one cuff and −0.66 [−1.23, −0.09] mmHg with five) and better than simply carrying the cuff value forward (−0.87 [−1.43, −0.31] mmHg with five cuffs). The Bland–Altman limits of agreement narrowed from ±32 mmHg without calibration to −26/+27 mmHg with five cuffs (Figure 3a,b), but the proportional error remained: high pressures were under- and low pressures over-estimated.

Table 2. Calibration-free estimation on unseen subjects (65 PPG features, 10 repeats; mean [95 % CI, subject-level (VitalDB: case-level) cluster bootstrap]; SD over repeats in the text). BHS: cumulative percentage of SBP errors within 5/10/15 mmHg.

Model	MIMIC-BP Test (229 Subjects)				VitalDB (17 Cases)	
	SBP MAE	SBP ME ± SD	DBP MAE	BHS (SBP)	SBP MAE	DBP MAE
Training mean	14.23 [13.40, 15.13]	−0.38 ± 17.69	9.35	21/41/59	14.06 [12.10, 16.93]	8.49
RF	13.03 [12.30, 13.81]	0.02 ± 16.41	8.24	23/46/65	14.23 [12.46, 16.51]	8.36
XGBoost	12.99 [12.25, 13.79]	−0.20 ± 16.39	8.21	24/46/65	14.21 [12.35, 16.70]	9.25
LightGBM	12.98 [12.25, 13.76]	−0.22 ± 16.37	8.22	24/46/65	14.09 [12.27, 16.70]	8.80
SVR	13.25 [12.44, 14.04]	−0.93 ± 16.79	8.35	24/46/65	14.57 [12.74, 16.78]	8.68
MLP	13.09 [12.37, 13.94]	−0.98 ± 16.54	8.27	24/46/65	14.87 [12.91, 17.30]	9.36
Sparse GPR	13.12 [12.38, 13.85]	−0.42 ± 16.53	8.28	24/46/65	14.77 [12.70, 16.96]	9.76

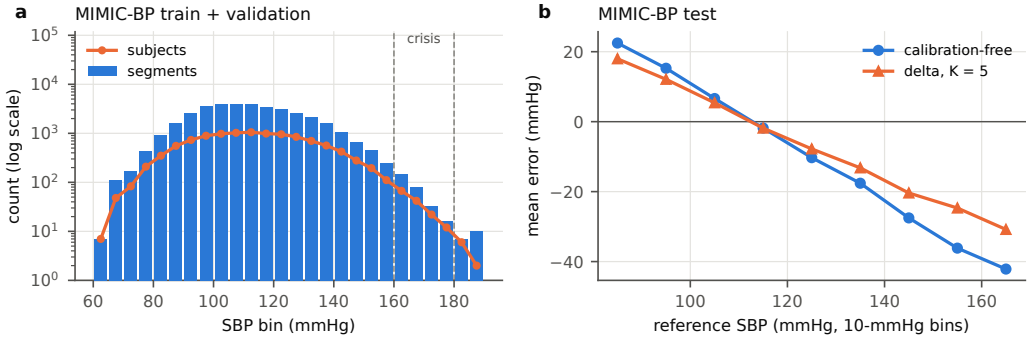


Figure 2. Distribution of blood pressure and its effect on the estimation error. (**a**) Number of segments (bars) and subjects (line) per 5 mmHg SBP bin in the MIMIC-BP training and validation pool (log scale); dashed lines mark 160 and 180 mmHg. (**b**) Mean error per 10 mmHg bin of the reference SBP on the MIMIC-BP test set for the calibration-free model and for delta calibration with five cuffs ($K = 5$); estimates averaged over the ten repeats, bins with fewer than 10 segments omitted. Without calibration, high pressures are under-estimated and low pressures over-estimated.

Part of the benefit of additional cuffs was due to their recency rather than their number (Table 3). On MIMIC-BP, a single cuff at the fifth segment gave an SBP MAE of 11.05 mmHg, 0.79 [0.21, 1.38] mmHg lower than a single cuff at the first segment (11.84 mmHg). With the most recent cuff fixed at the fifth segment, adding earlier cuffs gave SBP MAEs of 10.87 ($\{4, 5\}$), 10.76 ($\{3, 4, 5\}$) and 10.80 mmHg ($\{1, \ldots, 5\}$) compared with 11.05 mmHg ($\{5\}$); the change from $\{5\}$ to $\{1, \ldots, 5\}$ was −0.26 mmHg (95 % CI −0.56 to +0.04), providing no clear evidence of an additional SBP benefit, whereas the DBP MAE decreased from 6.31 to 5.98 mmHg (−0.33 [−0.56, −0.10] mmHg). In this design, about three quarters (0.79 of 1.04 mmHg) of the SBP improvement from one to five cuffs was therefore attributable to the shorter time since the last cuff, and the number of cuffs mattered more for DBP and for the constant method, which averages the cuff readings directly (14.10 with $\{5\}$ vs. 11.67 mmHg with $\{1, \ldots, 5\}$). On VitalDB, the five segments span only the first minute of the recording, so recency did not differ between the sets; there, the SBP differences between the sets were not significant, and five cuffs improved DBP over the fifth cuff alone (−1.08 [−1.95, −0.22] mmHg; Table 3). This agrees with the dominant role of the recalibration interval (Section 3.5).

Table 3. Number versus recency of the calibration cuffs (MAE in mmHg [95 % CI], cluster bootstrap; evaluation on the segments after the fifth usable segment). Positions refer to the usable segments of each subject. The models were re-trained with identical settings (ten repeats) and bootstrapped separately for this analysis; small differences from the calibrated-estimation results reported below reflect these separate runs.

Cuff Positions	MIMIC-BP Test (229 Subjects)			VitalDB (17 Cases)		
	Delta SBP	Delta DBP	Constant SBP	Delta SBP	Delta DBP	Constant SBP
{1}	11.84 [11.23, 12.51]	6.84 [6.36, 7.35]	14.61	14.29 [11.73, 17.21]	7.92 [6.53, 9.75]	24.47
{1, 2}	11.41 [10.88, 11.99]	6.49 [6.08, 6.94]	13.16	15.16 [12.38, 18.38]	7.96 [6.54, 9.79]	25.86
{1, 2, 3}	11.11 [10.59, 11.65]	6.27 [5.88, 6.69]	12.23	15.07 [12.21, 18.60]	7.72 [6.56, 8.95]	25.58
{5}	11.05 [10.51, 11.62]	6.31 [5.91, 6.77]	14.10	16.34 [12.16, 22.29]	8.50 [7.24, 10.03]	27.08
{4, 5}	10.87 [10.35, 11.42]	6.00 [5.63, 6.44]	12.84	15.91 [11.90, 21.45]	8.04 [6.74, 9.66]	25.89
{3, 4, 5}	10.76 [10.25, 11.30]	6.00 [5.62, 6.44]	12.14	15.75 [11.80, 21.21]	7.82 [6.41, 9.61]	25.37
{1, … , 5}	10.80 [10.30, 11.31]	5.98 [5.63, 6.38]	11.67	15.11 [11.81, 19.44]	7.42 [6.22, 8.78]	24.37
{1, … , 5} − {5}	−0.26 [−0.56, +0.04]	−0.33 [−0.56, −0.10]		−1.23 [−3.31, +0.29]	−1.08 [−1.95, −0.22]	
{5} − {1}	−0.79 [−1.38, −0.21]	−0.53 [−0.97, −0.10]		+2.05 [−2.31, +7.97]	+0.57 [−1.77, +2.63]	

The limitation is quantified in Figure 3c,d. Constant and offset calibration did not follow changes (tracking slope 0 and 0.07 on MIMIC-BP), whereas the delta model followed 38 % of the true change with one cuff on MIMIC-BP and 52 % on VitalDB; changes larger than 40 mmHg were under-estimated by about 30 mmHg. New-onset hypertensive segments (SBP ≥160 mmHg in subjects calibrated below 160 mmHg) were nevertheless ranked well, with an AUC of 0.85 [0.74, 0.96] with one cuff and 0.94 [0.85, 0.99] with five cuffs. These AUCs rest on few events: 30 positive segments from 9 subjects among 5324 segments of 227 subjects with one cuff, and 33 positive segments from 11 subjects among 5369 segments of 229 subjects with five cuffs. Because the estimates under-estimated high pressures, no positive segment reached an estimate of 160 mmHg (sensitivity 0 and specificity 1.00 at that threshold); the AUC reflects the correct ordering of the estimates, not their absolute level. A lower alert threshold would be required; for illustration only, the threshold that maximized Youden's index on the test set itself (122 mmHg with five cuffs) gave a sensitivity of 0.94 and a specificity of 0.85, which is optimistic because the threshold was not chosen based on independent data. On VitalDB, the AUC was 0.56 [0.36, 0.72] with one cuff (477 positive segments from 11 cases). Because the pooled segment-level AUC mixes differences between and within subjects, we also computed a within-subject AUC (mean over the subjects that had both positive and negative segments) and a subject-level AUC (maximum estimate of each subject versus whether the subject ever reached 160 mmHg), and the segment- and subject-level performance of fixed alert thresholds (Table 4). Carrying the cuff value forward reached a pooled AUC of 0.81, but its within-subject AUC was 0.50 by construction, confirming that its pooled AUC reflects only between-subject differences in the calibration value. The delta model reached within-subject AUCs of 0.77 [0.63, 0.90] with one cuff and 0.83 [0.74, 0.92] with five cuffs on MIMIC-BP and 0.73 (one cuff) and 0.67 (recalibration every 15 min) on VitalDB, so it did rank the segments of the same subject. As an alarm, however, it performed poorly: no estimate reached 160 mmHg, and with a lower threshold of 140 mmHg (fixed for this revision, not optimized) the segment-level sensitivity was 0.15–0.34, the precision 0.18–0.36 and the F1-score 0.19–0.26 on MIMIC-BP, with 0.2–0.9 false alarms per hour of monitoring; at the subject level, 55 % (five cuffs) to 69 % (recalibration) of the 11–13 subjects with a new-onset event were flagged at least once. With 9–13 positive subjects per configuration, all of these estimates are uncertain.

On VitalDB, a single intra-operative cuff did not improve SBP over the calibration-free model (14.27 vs. 14.18 mmHg) because the pressure moved far from its initial value during surgery, although DBP improved (7.93 vs. 9.24 mmHg)

Table 4. Detection of new-onset SBP ≥160 mmHg in subjects calibrated below 160 mmHg [95 % CI, cluster bootstrap]. Alarm: estimate ≥140 mmHg. False alarms per hour: segments with reference SBP <160 mmHg and estimated SBP ≥140 mmHg, as a rate per hour of cumulative monitored time among segments with reference SBP <160 mmHg (MIMIC-BP 30 s, VitalDB 10 s segments).

	Configuration	Pos. Subj./Seg.	AUC Pooled	AUC Within-Subject	AUC Subject	Sens./Spec. (Segment)	F1 (Segment)	False Alarms/h
MIMIC-BP	1 cuff, constant	9/30	0.81 [0.65, 0.93]	0.50	0.79 [0.60, 0.94]	0.27/0.939	0.04	7.4 [3.8, 11.6]
	1 cuff, delta	9/30	0.85 [0.74, 0.96]	0.77 [0.63, 0.90]	0.83 [0.64, 0.96]	0.20/0.995	0.19	0.6 [0.0, 1.5]
	5 cuffs, delta	11/33	0.94 [0.85, 0.99]	0.83 [0.74, 0.92]	0.92 [0.77, 1.00]	0.15/0.998	0.21	0.2 [0.0, 0.4]
	Recal. 5 segments	13/29	0.93 [0.89, 0.97]	0.80 [0.73, 0.88]	0.95 [0.90, 0.99]	0.34/0.993	0.26	0.9 [0.4, 1.6]
VitalDB	1 cuff, constant	11/477	0.42 [0.23, 0.65]	0.50	0.73 [0.42, 0.96]	0.07/0.864	0.02	49 [0, 125]
	1 cuff, delta	11/477	0.56 [0.36, 0.72]	0.73 [0.64, 0.83]	0.53 [0.21, 0.85]	0.00/0.980	0.00	7.1 [0.0, 22.4]
	Recal. 15 min	12/447	0.72 [0.47, 0.82]	0.67 [0.51, 0.80]	0.77 [0.47, 0.98]	0.02/0.998	0.04	0.8 [0.1, 1.9]

The mean errors of all calibrated configurations were close to zero (at most 1.3 mmHg in absolute value; Table 5), so the error was dominated by its SD, and the RMSE equalled the SD within 0.1 mmHg (e.g., 13.52 mmHg SBP with five cuffs). Referenced to the AAMI/ESH/ISO criterion (ME ≤ 5 mmHg and SD ≤ 8 mmHg), SBP did not meet the SD limit in any configuration (12.4–18.9 mmHg), whereas DBP met it with five cuffs (-0.21 ± 7.99 mmHg) and with periodic recalibration (MIMIC-BP every 5 segments -0.35 ± 7.26 mmHg; VitalDB every 5 and 15 min $+0.98 \pm 6.29$ and $+1.02 \pm 7.39$ mmHg). The BHS cumulative percentages within 5, 10 and 15 mmHg were 28, 54 and 74 % for SBP and 54, 82 and 94 % for DBP with five cuffs, and 37, 63 and 79 % and 64, 87 and 94 % with recalibration every 5 segments; SBP thus remained below grade C of the BHS protocol, while DBP reached grade B.

Table 5. Calibrated estimation (reference delta model, 10 repeats). MAE in mmHg with subject-level (VitalDB: case-level) bootstrap 95 % CI; ME ± SD, mean error and SD of the error in mmHg (RMSE equals the SD within 0.1 mmHg). New-onset AUC: detection of SBP ≥160 mmHg in subjects calibrated below 160 mmHg.

Configuration	SBP MAE	SBP ME ± SD	DBP MAE	DBP ME ± SD	Onset AUC
MIMIC-BP test (229 subjects)					
No calibration	12.99 [12.25, 13.77]	-0.17 ± 16.30	8.19 [7.63, 8.73]	-0.76 ± 10.54	0.66
1 cuff, constant	14.61 [13.66, 15.64]	$+0.55 \pm 18.95$	8.72 [8.02, 9.48]	$+0.36 \pm 11.67$	0.81
1 cuff, delta	11.84 [11.21, 12.48]	$+0.44 \pm 14.93$	6.84 [6.35, 7.35]	$+0.06 \pm 9.10$	0.85 [0.74, 0.96]
3 cuffs, delta	11.11 [10.57, 11.66]	$+0.23 \pm 13.91$	6.27 [5.85, 6.70]	-0.09 ± 8.30	0.92
5 cuffs, delta	10.79 [10.29, 11.30]	$+0.17 \pm 13.52$	5.98 [5.59, 6.37]	-0.21 ± 7.99	0.94 [0.85, 0.99]
Recal. 5 segments	9.45 [9.00, 9.89]	-0.38 ± 12.53	5.04 [4.72, 5.36]	-0.35 ± 7.26	0.93
Recal. 10 segments	10.41 [9.89, 10.99]	-0.10 ± 13.68	5.64 [5.29, 6.03]	-0.36 ± 7.95	0.88
VitalDB (17 cases)					
No calibration	14.18 [12.41, 16.61]	$+0.73 \pm 18.45$	9.24 [7.75, 10.76]	$+3.54 \pm 10.64$	0.42
1 cuff, delta	14.27 [11.64, 17.42]	$+0.46 \pm 18.92$	7.93 [6.52, 9.75]	$+0.47 \pm 10.36$	0.56
Recal. 5 min	8.66 [7.58, 10.01]	-0.45 ± 12.38	4.44 [4.11, 4.81]	$+0.98 \pm 6.29$	0.84
Recal. 15 min	10.32 [8.92, 12.32]	-0.56 ± 14.61	5.27 [4.81, 5.80]	$+1.02 \pm 7.39$	0.72
Recal. 30 min	11.13 [9.69, 13.10]	-0.44 ± 15.38	5.87 [5.37, 6.51]	$+1.32 \pm 7.92$	0.73

3.3. Errors by Blood Pressure Level and During Large Changes

Table 6 separates the errors by the level of the reference SBP. In the normotensive range (90–140 mmHg), which contained 84 % of the MIMIC-BP test segments, the delta model with one cuff had an SBP MAE of 9.77 mmHg, and recalibration every 5 segments reduced it to 7.74 mmHg. Outside this range, the errors were large and systematic: below 90 mmHg, the estimates were too high (mean error +21.1 mmHg with one cuff) and, at 140–160 and ≥160 mmHg, too low (−22.2 and −37.2 mmHg). Calibration and recalibration reduced but did not remove this bias, and on VitalDB the errors at ≥160 mmHg remained about 47–55 mmHg. The results above 160 mmHg in MIMIC-BP rest on only 33 segments from 11 subjects.

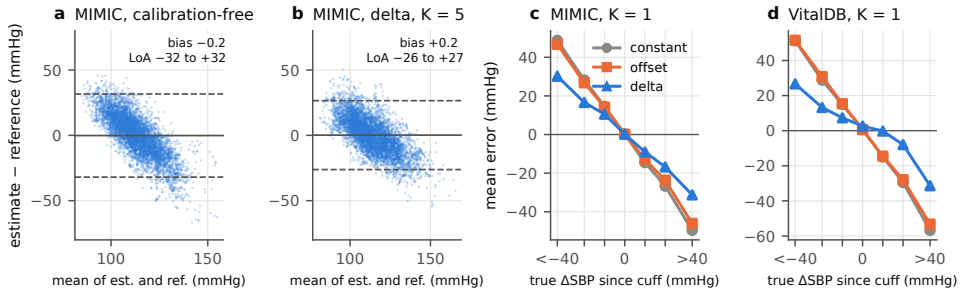


Figure 3. Calibration and tracking of SBP changes. (**a**,**b**) Bland–Altman plots on the MIMIC-BP test set for the calibration-free model and for delta calibration with $K = 5$ (estimates averaged over the ten repeats); solid line, mean difference; dashed lines, 95 % limits of agreement. (**c**,**d**) Mean error versus the true SBP change since the calibration cuff ($K = 1$) on MIMIC-BP (**c**) and VitalDB (**d**) for constant, offset and delta calibration; bins of the change: <−40, −40 to −20, −20 to −10, −10 to 10, 10 to 20, 20 to 40 and >40 mmHg; bins with fewer than 10 segments omitted. A negative slope indicates that the change is under-estimated.

Table 7 shows the errors separately for rises and falls of SBP since the calibration cuff. With one cuff on MIMIC-BP, the model reproduced 37 % of rises of more than 40 mmHg and 38 % of falls of more than 40 mmHg, with MAEs of about 30 mmHg; carrying the cuff value forward reproduced none and gave MAEs of about 49 mmHg. On VitalDB, 45 % of rises and 48 % of falls larger than 40 mmHg were reproduced. The under-estimation was therefore symmetric for rises and falls. For changes exceeding 40 mmHg, the delta model reduced MAE by approximately 37–48 % relative to carrying the cuff value forward, while still substantially underestimating the changes.

Table 6. SBP MAE and mean error (ME) by reference SBP level (delta model; mmHg; 95 % CI by cluster bootstrap). *n*: segments (subjects or cases).

	Reference SBP	*n*	1 Cuff: MAE/ME	5 Cuffs: MAE/ME	Recalibration [a]: MAE/ME
MIMIC-BP	<90	461 (91)	21.1 [18.7, 23.6]/+21.1	20.0 [17.8, 22.2]/+20.0	16.0 [14.3, 17.9]/+15.9
	90–140	4522 (227)	9.8 [9.3, 10.3]/+0.4	8.8 [8.4, 9.2]/+0.1	7.7 [7.4, 8.1]/−0.4
	140–160	353 (81)	22.3 [19.9, 24.5]/−22.2	21.6 [20.0, 23.4]/−21.6	18.9 [17.3, 20.8]/−18.8
	≥160	33 (11)	37.2 [25.2, 43.6]/−37.2	34.8 [29.1, 39.3]/−34.8	28.6 [19.9, 34.6]/−28.6
VitalDB	<90	1107 (16)	27.3 [22.3, 35.5]/+27.3	27.4 [23.1, 34.2]/+27.4	22.0 [18.8, 25.5]/+22.0
	90–140	20,810 (17)	10.9 [9.1, 13.4]/+3.0	12.3 [9.5, 16.9]/+4.4	7.6 [7.0, 8.3]/+1.1
	140–160	1682 (16)	29.4 [24.4, 33.4]/−29.4	26.7 [19.4, 32.4]/−26.4	20.2 [17.7, 22.7]/−20.2
	≥160	573 (12)	54.7 [43.8, 59.8]/−54.7	52.2 [38.9, 58.0]/−52.2	46.7 [39.1, 52.2]/−46.7

[a] MIMIC-BP every 5 segments; VitalDB every 15 min. Recalibration rows differ slightly in number (see Section 2.5).

Table 7. SBP errors during rises and falls since the calibration cuff (one cuff; mmHg). Captured: mean estimated change divided by mean true change [95 % CI].

	True Change	*n*	True/Estimated Change	Captured (%)	MAE, Delta [95 % CI]	MAE, Carried-Forward Cuff
MIMIC-BP	fall >40	124 (39)	−48.7/−18.6	38 [33, 44]	30.2 [26.5, 33.2]	48.7
	fall 20–40	600 (101)	−28.1/−11.6	41 [34, 48]	17.0 [15.4, 18.7]	28.1
	within ±20	3962 (227)	−0.2/+0.2	–	9.2 [8.6, 9.8]	8.8
	rise 20–40	571 (113)	+26.8/+9.9	37 [29, 44]	17.1 [15.3, 19.1]	26.8
	rise >40	112 (31)	+49.7/+18.5	37 [26, 46]	31.2 [27.1, 35.4]	49.7
VitalDB	fall >40	1579 (7)	−51.3/−24.6	48 [29, 70]	26.8 [16.3, 35.9]	51.3
	fall 20–40	3898 (11)	−28.7/−15.5	54 [28, 75]	14.6 [9.7, 21.3]	28.8
	within ±20	11,318 (17)	−0.8/+2.4	–	9.6 [8.7, 10.9]	9.8
	rise 20–40	4817 (16)	+29.5/+21.6	73 [43, 90]	10.4 [6.4, 18.9]	29.5
	rise >40	2560 (15)	+56.8/+25.4	45 [25, 61]	31.4 [22.2, 43.4]	56.8

3.4. Effect of Cuff Measurement Error

Adding a random error to the simulated cuff readings increased the error in all calibration methods, but much less for the delta model than for carrying the cuff value forward or for offset calibration (Table 8). With one cuff on MIMIC-BP, a cuff error with SD 5 and 10 mmHg increased the delta-model SBP MAE from 11.84 to 11.96 and 12.40 mmHg, and the DBP MAE from 6.84 to 7.37 and 8.65 mmHg; the corresponding increases for the carried-forward cuff were 0.49 and 1.99 mmHg (SBP). With five cuffs, the effect was smaller (SBP 10.80 to 10.95 mmHg at SD 10 mmHg), because independent errors average out. Recalibration was the most sensitive to cuff error, because every new cuff brings a new error: on VitalDB with a cuff every 15 min, the SBP MAE increased from 10.37 to 10.62 and 11.29 mmHg and the DBP MAE from 5.31 to 6.18 and 7.90 mmHg. A constant bias of +5 mmHg increased the delta-model SBP MAE by a further 0.27–0.58 mmHg across all simulated configurations (0.27–0.36 mmHg for the configurations reported in the Table 8 footnote). The ranking of the methods and the benefit of recalibration were unchanged. However, DBP, which met the 8 mmHg SD criterion of the standard with an error-free cuff and recalibration (SD 7.3 mmHg on MIMIC-BP, 7.4 mmHg on VitalDB at 15 min), no longer met it with a cuff error of SD 5 mmHg (8.1 and 8.2 mmHg) or 10 mmHg (10.1 and 10.1 mmHg; Table 8, DBP error SD); only recalibration nominally every 5 min on VitalDB still met it with a cuff error of SD 5 mmHg (7.31 mmHg). The accuracy of the cuff used for calibration therefore directly limits the attainable accuracy of a calibrated device.

Table 8. Effect of simulated cuff measurement error (independent Gaussian error added to every cuff reading; zero mean; SD 0, 5 or 10 mmHg). MAE in mmHg, mean over ten repeats and five error draws; errors computed against the invasive reference.

Data	Configuration	SBP MAE			DBP MAE			DBP Error SD
		SD 0	SD 5	SD 10	SD 0	SD 5	SD 10	SD 0/5/10
MIMIC-BP	1 cuff, constant	14.61	15.10	16.60	8.72	9.66	11.95	11.67/12.66/15.28
	1 cuff, offset	14.38	14.88	16.39	8.55	9.49	11.81	11.35/12.38/15.06
	1 cuff, delta	11.84	11.96	12.40	6.84	7.37	8.65	9.11/9.65/11.09
	5 cuffs, delta	10.80	10.83	10.95	5.98	6.11	6.49	7.99/8.13/8.53
	Recal. 5 segments, delta	9.45	9.80	10.67	5.04	6.01	7.83	7.26/8.10/10.07
	Recal. 5 segments, offset	10.43	11.39	13.57	5.97	7.43	10.31	8.81/10.11/13.29
VitalDB	1 cuff, delta	14.29	14.47	14.82	7.92	8.56	9.76	10.36/11.01/12.35
	Recal. 5 min, delta	8.73	9.07	9.92	4.48	5.53	7.50	6.30/7.31/9.56
	Recal. 15 min, delta	10.37	10.62	11.29	5.31	6.18	7.90	7.42/8.23/10.13
	Recal. 30 min, delta	11.21	11.40	11.98	5.90	6.59	8.11	7.94/8.63/10.34
	Recal. 15 min, offset	12.30	13.16	15.18	6.33	7.72	10.56	8.98/10.33/13.52

The SD 0 results were obtained from the re-trained models and differed from the corresponding results in Table 5 by at most 0.08 mmHg. DBP error SD: SD of the DBP error (mmHg) at cuff error SD 0/5/10. With an additional bias of +5 mmHg, the delta-model SBP MAE was 12.23/12.72 (1 cuff, SD 5/10) and 10.13/11.03 (recal. 5 segments).

3.5. Recalibration Interval and Cuff Policy

Accuracy was governed by the time since the last cuff. With a cuff every 5 segments, the delta model reached 9.45 [9.00, 9.89] mmHg SBP and 5.04 mmHg DBP on MIMIC-BP; every 10 segments, 10.41 mmHg. On VitalDB, recalibration every 5, 15 and 30 min gave SBP MAEs of 8.66, 10.32 and 11.13 mmHg with the delta model; offset calibration was equivalent at 5 min (8.36 mmHg) but degraded faster (12.29 and 14.49 mmHg at 15 and 30 min). Figure 4 shows that the MAE decreased smoothly with the cuff rate, from 12.3 mmHg at one cuff per hour to 8.0 mmHg at one cuff every 3 min (VitalDB). Event-triggered cuffs, requested when the predicted change exceeded a threshold, did not beat periodic cuffs at the same rate: under the pre-specified comparison, no trigger or hybrid policy was selected on the MIMIC-BP validation subjects or in any of the 17 VitalDB leave-one-case-out folds, and at low cuff rates the best trigger policy was worse than periodic recalibration (VitalDB, one cuff per 15 min: 10.95 vs. 10.33 mmHg). Because the external set comprised only 17 cases, the errors are reported for every case in Table A2. With recalibration every 15 min, the

case-level SBP MAE ranged from 6.5 to 19.9 mmHg (median 9.9 mmHg) and was lower than without calibration in all 17 cases; with a single cuff, it was lower in only 8 of 17 cases. The case-level error was closely related to the variability in the pressure during the case (correlation between the case MAE with 15 min recalibration and the within-case SD of the reference SBP, $r = 0.94$); cases 0001 and 0010 had within-case SBP SDs above 27 mmHg and large estimation errors. The case-level bootstrap CIs in the tables reflect this heterogeneity. A subject-specific gain on the predicted change reduced the MAE only marginally (−0.14 [−0.23, −0.06] mmHg on MIMIC-BP at one cuff per 5 segments).

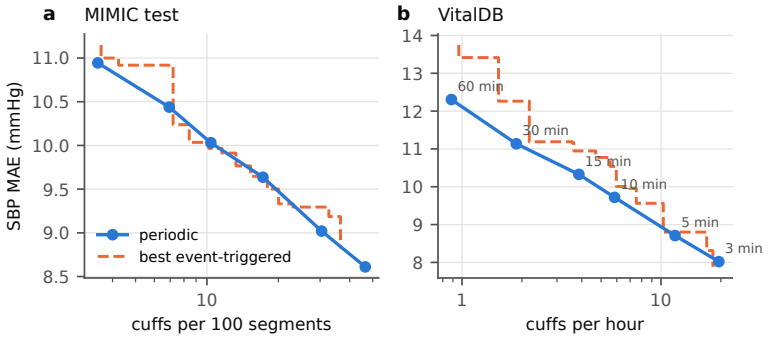


Figure 4. SBP MAE of the delta model versus the number of cuff readings for periodic recalibration (solid) and for the best event-triggered or hybrid policy at the same or a lower cuff rate (dashed; lower envelope of all configurations), averaged over the ten repeats. (**a**) MIMIC-BP test set, cuffs per 100 segments (30 s segments); (**b**) VitalDB, cuffs per hour, with the recalibration interval annotated. Logarithmic horizontal axes.

3.6. Lightweight Models

Model size could be reduced by more than an order of magnitude with little loss of accuracy (Table 9, Figure 5). The pre-specified rule selected a small gradient-boosted model (depth 3, up to 200 trees) with the 20 most important features, which needs about 6000 tree nodes (an estimated 47 KB) and about 1200 comparisons per estimate (2 targets × 200 trees of depth 3, i.e., three comparisons per tree), compared with 190,000 nodes (1.5 MB) for the reference. Its SBP MAE with one cuff was 12.06 mmHg, a difference of +0.22 [+0.07, +0.37] mmHg from the reference; the upper confidence bound was below the pre-specified non-inferiority margin of 0.5 mmHg, although the loss was statistically detectable. The differences were +0.28 [+0.16, +0.40] mmHg with five cuffs and +0.40 [+0.30, +0.51] mmHg with recalibration every 5 segments on MIMIC-BP, and +0.38 [−0.10, +0.97] mmHg with one cuff and +0.19 [+0.03, +0.39] mmHg with 15 min recalibration on VitalDB (paired case-level bootstrap). The random forest was the least suitable model: 300 trees required 3.8 million nodes (about 30 MB) and 20 ms per estimate without being more accurate (12.19 mmHg). Ridge regression (266 parameters) and a small MLP (4322 parameters) were competitive on MIMIC-BP (12.15 and 12.18 mmHg) but failed on the external VitalDB cases with a single cuff (19.63 and 21.69 mmHg vs. 14.27 mmHg), and remained worse after 15 min recalibration (11.24 and 13.92 vs. 10.32 mmHg). The model size does not include the pre-processing and feature extraction. The XGB-small model consists of two ensembles (SBP and DBP) of 200 trees of depth 3 (all 200 trees were retained by early stopping), i.e., at most 15 nodes per tree and 5987 nodes in total; at 8 bytes per node (feature index, threshold, child offset), this gives 47 KB, and one estimate requires exactly three comparisons per tree, i.e., 1200 comparisons (the 1600 given in the original submission was an upper bound computed as $\lceil \log_2(\text{nodes}) \rceil$ per tree). The complete pipeline additionally needs a buffer of one 10 s segment (1250 samples, 5 KB in single precision), a third-order band-pass filter applied forward and backward (about 3×10^4 multiply–adds), first and second derivatives, beat

delineation and per-beat feature computation, each linear in the number of samples, plus a multitaper spectrum per beat for one feature and a two-Gaussian least-squares fit per beat for five morphology features. To quantify the cost, we timed a Python re-implementation of the complete feature extraction on a desktop CPU (single thread, 200 MIMIC-BP segments): the 49 published and standard features took 90 ms per 10 s segment (beat detection 0.7 ms), the 16 morphology descriptors 165 ms with the Gaussian fit and 3.1 ms without it, and one model evaluation 0.4 ms; the complete pipeline thus took about 255 ms per 10 s segment (2.6 % of real time), or about 95 ms without the Gaussian fit, and the peak Python heap during the extraction of one segment was 485 KB, most of it interpreter overhead. Removing the five Gaussian features from the small model changed its SBP MAE by −0.03 mmHg on MIMIC-BP with one cuff (11.98 vs. 12.01 mmHg, ten repeats) and by +0.10 mmHg on VitalDB with 15 min recalibration (10.51 vs. 10.41 mmHg), so the fit can be omitted on a device at a negligible cost in accuracy. These desktop Python measurements do not establish runtime or memory requirements on embedded hardware; on-device latency, peak RAM, and energy consumption remain to be measured. The inference-latency figure above (0.4 ms) was measured on an XGBoost pair with the same architecture as the deployed small model (132 input features, 200 trees, depth 3) rather than the deployed model's own weights, because per-sample inference time for an already-grown tree ensemble is determined by the number and depth of the trees actually walked, not by the specific split thresholds or leaf values learned during training; the two are therefore equivalent for this measurement, and this does not affect the model-size or operation counts above, which come from the deployed model itself. The ten features selected in all ten repeats were dominated by second-derivative (acceleration PPG) and pulse-shape descriptors: DT, the height and width of the second-derivative wave, the height ratio of its *d* and *e* waves to the *a* wave, kurtosis, Hjorth mobility, the relative dicrotic-notch height, the 50 % up-stroke time and the width of the first Gaussian component.

Table 9. Delta-calibration models of different size (10 seeds, SBP MAE in mmHg). Memory is estimated for a flat on-device representation (8 bytes per tree node, 4 bytes per weight); operations are comparisons (trees; upper bound $\sum_{\text{trees}} \lceil \log_2(\text{nodes}) \rceil$, for XGB-small exactly 3 per tree, i.e., 1200) or multiply–adds per estimate. * Selected by the pre-specified rule. [b] Recalibration every 5 segments.

Model	Feat.	Params	Memory (KB)	Ops	MIMIC-BP				VitalDB	
					Val. $K = 1$	Test $K = 1$	Test $K = 5$	Test Recal. [b]	$K = 1$	Recal. 15 min
XGB (reference)	65	189,505	1481	11,664	11.81	11.84	10.79	9.45	14.27	10.32
XGB	20	167,467	1308	10,411	11.92	11.93	10.92	9.54	14.58	10.42
LGBM-small	65	10,831	85	1494	11.96	11.95	10.92	9.61	14.44	10.40
XGB-small	65	5988	47	1200	12.07	12.02	11.02	9.79	14.65	10.41
XGB-small *	20	5987	47	1200	12.10	12.06	11.07	9.85	14.66	10.51
RF	65	3,805,398	29,730	11,811	12.13	12.19	11.04	9.53	14.53	10.72
RF-small	65	19,747	154	400	12.50	12.41	11.19	9.80	15.64	10.61
Ridge	65	266	1	264	12.38	12.15	11.29	10.32	19.63	11.24
MLP-small	65	4322	17	4288	12.20	12.18	11.07	10.23	21.69	13.92

3.7. State Tracking and Residual Dynamics

Kalman filtering of the delta estimates between cuffs added little: on VitalDB, it reduced the SBP MAE by 0.2–0.5 mmHg (e.g., 10.23 to 9.73 mmHg with 15 min recalibration, 3 repeats), and on MIMIC-BP by at most 0.13 mmHg; the unscented filter with bias and gain states performed like the linear filter. The reason is the structure of the residual: on VitalDB, the autocorrelation of the delta-model SBP error was 0.96 at 10 s, 0.66 at 5 min, 0.49 at 15 min and 0.32 at 60 min, and 84 % of its variance was slow (5 min running mean). The error is therefore a slowly drifting bias that only a new cuff can remove, which explains why accuracy depends on the recalibration interval rather than on filtering or model complexity.

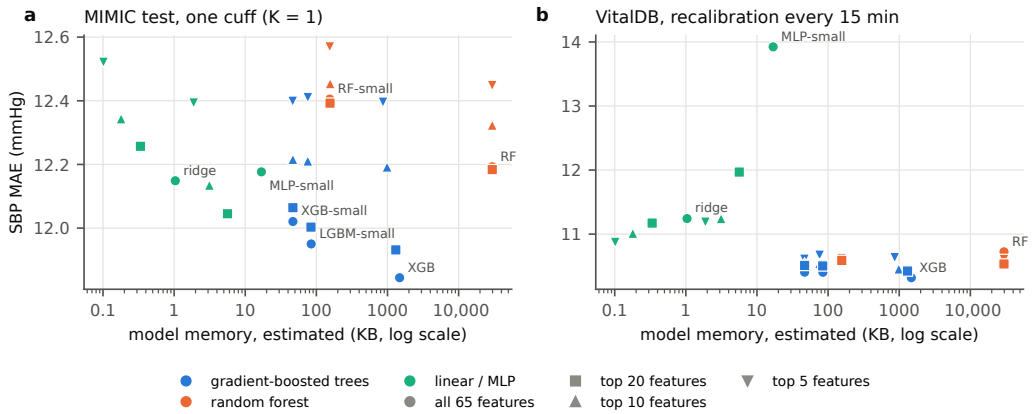


Figure 5. SBP MAE versus estimated model memory for all variants of the delta model (8 bytes per tree node, 4 bytes per weight; logarithmic axis), averaged over the ten repeats. (**a**) MIMIC-BP test set with one calibration cuff; (**b**) VitalDB with recalibration every 15 min. Color, model family; marker, number of input features (all 65 or the 20, 10 or 5 most important).

4. Discussion

4.1. Calibration Is Required for Unseen Users

On subjects not seen during training, six regressors of very different complexity converged to the same SBP error of about 13 mmHg, only 1.2 mmHg below a constant predictor, and on an external database they did not beat the constant predictor at all. That the choice of model did not matter indicates that the limitation lies in the input rather than in the learning algorithm: hand-crafted PPG features describe the pulse shape of a person more than the absolute pressure, so a model trained on other people regresses its estimates towards the population mean (Figure 2b). This is consistent with the recent scientific statement of the American Heart Association, which notes that many cuffless devices need an initial and repeated cuff calibration [21], and with reports that calibration-free accuracy is overestimated when segments of the same subject appear in training and test sets [18]. For a smart-device implementation, the practical question is therefore not whether to calibrate, but how many cuffs, how often, and with how large a model.

4.2. How Many Cuffs and How Often

With the delta model, one cuff reduced the SBP error to 11.8 mmHg and five cuffs to 10.8 mmHg, and the benefit over simpler offset calibration was significant for every number of cuffs. However, most of the SBP benefit of five cuffs came from the shorter time since the last of them; with the last cuff fixed, additional earlier cuffs changed SBP by −0.26 mmHg (95 % CI −0.56 to +0.04) and improved DBP by 0.33 mmHg (Table 3). For a device, a recent cuff is therefore worth more than several older ones. The largest gain, however, came from recalibration. Because the residual of the calibrated model is a slowly drifting bias (autocorrelation 0.49 after 15 min on VitalDB), accuracy decays with the time since the last cuff, and neither Kalman filtering nor event-triggered cuffs could substitute for new absolute information. Periodic recalibration was the best policy at every cuff budget we examined. The under-estimation of large changes is part of the reason. An event-triggered policy requests a cuff when the predicted change exceeds a threshold, but the model reproduced only 37–48 % of changes larger than 40 mmHg (Table 7); a true rise of 40 mmHg therefore appears as a predicted rise of about 15–20 mmHg, so a trigger set at 20 mmHg fires late or not at all, and a lower threshold also fires on the larger number of small, spurious predicted changes. In addition, the residual error is a slow drift (Section 3.7) that is not visible in the

predicted change at all. Both properties favor cuffs at fixed intervals. This loss of accuracy with time since calibration is precisely what current validation protocols do not examine, since they test devices immediately after calibration [21]. The resulting curve (Figure 4) quantifies it and can be read as a design rule: on VitalDB, recalibration after every 90 retained 10 s segments—nominally, 15 min—yielded an SBP MAE of 10.32 mmHg, and recalibration nominally every 5 min, 8.66 mmHg. This quantifies the trade-off between recalibration burden and error in the surgical dataset; it does not establish a suitable recalibration interval for ambulatory or home use, which has to be determined on ambulatory data. The intervals of 5, 15 and 30 min examined here are practical only where an automated cuff is already in place, for example, in perioperative or intensive-care monitoring, where non-invasive cuffs are routinely cycled automatically at intervals of a few minutes; there, the PPG estimate would bridge the cuff readings. For a consumer device, a cuff every 15 min is not acceptable, and our data do not determine which interval would be adequate in daily life; they do imply that the stated accuracy of a calibrated device applies only within the interval for which it was validated, so that the time since the last cuff should be reported together with the estimate. The curve in Figure 4 quantifies this trade-off; it does not prescribe a schedule for daily use.

4.3. Small Models Suffice

Reducing the delta model from about 190,000 to 6000 tree nodes (from an estimated 1.5 MB to 47 KB) cost about 0.2 mmHg, which is small compared with the 2–3 mmHg gained by recalibrating every 5 segments instead of once. Accuracy was therefore limited by the information available to the model rather than by its capacity, and a model of this size can be evaluated with about 1200 comparisons. This is within the budgets of microcontroller-class devices used for on-device machine learning, whose RAM ranges from 8 to 264 kB for ultra-constrained processors and from about 60 to 512 kB for standard Cortex-M4/M7 microcontrollers, with frozen model parameters typically held in flash memory [37]. Because physiological signals are person-specific and the data distribution changes after deployment, adaptation on the device is an active topic in this field [37]; the delta calibration studied here is a very light form of such adaptation, requiring only the stored features and value of the last cuff. Two findings are relevant for the choice of model family. First, the random forest was large and slow without being more accurate. Second, the linear model and the small neural network, although the smallest, failed on the external database, where the pressure range and the PPG acquisition differed; tree ensembles saturate outside the training range, whereas linear and smooth models extrapolate, which is a disadvantage when a device meets users and conditions not represented in its training data. The features retained by the small model were mainly derived from the second derivative of the PPG and from the pulse shape, which are robust to amplitude changes and can be computed from a single optical channel.

4.4. Limits of Trend Tracking

Calibration does not make the estimate follow large changes faithfully. The delta model reproduced 30–50 % of the change since the cuff and under-estimated changes larger than 40 mmHg by about 30 mmHg. This agrees with reports that cuffless devices often track pressure changes little better than a baseline that uses only the calibration cuff [21]; in our data, the delta model was significantly better than carrying the cuff forward, but only by 0.9 mmHg with five cuffs. Whether the gain over carrying the cuff value forward is practically relevant depends on the setting. With one cuff, it was 2.8 mmHg SBP and 1.9 mmHg DBP on MIMIC-BP, about one fifth of the error, and, for changes exceeding 40 mmHg, the delta model reduced MAE by approximately 37–48 % relative to carrying the

cuff value forward, while still substantially underestimating the changes (Table 7); with five cuffs, the gain was only 0.9 mmHg SBP. Relative to the remaining SBP error of 11–15 mmHg, these gains do not change the clinical usability of the SBP estimate, which remained far from the accuracy required of a measuring device (see below); they matter mainly for DBP and for detecting the direction of a change. Its predicted changes were nevertheless largely correctly ordered: with five cuffs, new-onset SBP ≥160 mmHg on MIMIC-BP was discriminated with a pooled AUC of 0.94, but sensitivity was zero at an estimated-SBP threshold of 160 mmHg. These results indicate the discrimination of new-onset high-SBP segments in MIMIC-BP, but do not establish a clinically usable alert. Lower alert thresholds require independent validation, particularly given the poor external alarm performance (VitalDB sensitivity 0.00–0.02 at 140 mmHg; Table 4). The estimate should not be used to report the magnitude of a large change without a cuff measurement. The number of new-onset events in the test set was small (30–33 positive segments from 9–11 subjects), so this estimate is uncertain.

4.5. Clinical Accuracy

The SBP results did not meet the AAMI/ESH/ISO accuracy criterion in any configuration: the SD of the SBP error was 12.4–18.9 mmHg against a limit of 8 mmHg, and the BHS grade remained below C. Only DBP met the SD criterion, and only with five cuffs or periodic recalibration. Cuff measurement error reduced the number of configurations meeting the DBP error-SD threshold; however, nominal 5 min recalibration on VitalDB still yielded an error SD of 7.31 mmHg with a cuff-error SD of 5 mmHg (Section 3.4, Table 8). The errors were largest exactly where accuracy matters most, at high and low pressures (Table 6). The approach studied here is therefore not a substitute for a cuff measurement; at best, it can provide a trend between cuff readings and prompt a cuff measurement when a change is suspected.

4.6. Comparison with Previous Work

Our errors are larger than many reported for MIMIC-derived data, for example, SBP MAE of 3.52 ± 5.10 mmHg with an MLP-Mixer on PPG and ECG [14], 3.91 ± 5.65 mmHg with a transformer and attention-gated recurrent units on PPG, ECG and pulse-transit features [16], 4.44 mmHg with a parallel CNN–transformer on PPG [17], and an SBP error of -0.17 ± 4.62 mmHg with a multimodal ECG–PPG network [15]. These studies differ from ours in the input (several use the ECG or the raw waveform), in the database version and segment selection (e.g., SBP restricted to 80–180 mmHg [15]), and, above all, in how training and test data are separated. In [15], the non-overlapping 10 s segments were randomly assigned to the training, validation and test sets, and in [16] the pooled samples of 1000 subjects were randomly shuffled before the split, so that segments of the same subject can appear in both training and test data. A model can then recognize the subject rather than estimate the pressure, and a public benchmark showed that such leakage produces over-optimistic results [18]. That benchmark proposed to express the error relative to a predictor returning the training mean (mean absolute scaled error) and reported values of about 89–93 % for support vector regression [18]; our calibration-free models reached 91 % of the mean-predictor error on MIMIC-BP (12.98 vs. 14.23 mmHg) and did not improve on it on VitalDB.

The study closest to ours in design is that of Chen et al. [13], who estimated the change in blood pressure from differences in PPG features relative to a baseline measurement and added the baseline cuff value, which corresponds to our delta calibration. In 90 healthy volunteers who were followed for three months, with subject-disjoint training and test sets, a random forest reached an RMSE of 7.15 mmHg for SBP and 5.30 mmHg for DBP [13]. Their

cohort covered SBP from 85 to 165 mmHg with changes of −25 to +39 mmHg from baseline, whereas our intensive-care and surgical data contain larger and faster changes, which the delta model under-estimates (Section 4.4). Both studies thus indicate that differential features with a cuff reference are a viable approach for unseen users. Our contribution is to quantify, on unseen subjects and on an external database, how the error in this approach depends on the number of cuffs, the time since the last cuff and the size of the model, which are the pieces of information needed to design a calibrated smart device.

4.7. Limitations

First, both databases contain finger PPG recorded in intensive care or during surgery, not wrist PPG in daily life, and the reference was invasive arterial pressure, which also served as the simulated cuff reading; the effect of cuff error was examined only with a simulated random error (Section 3.4), whereas a real cuff may also have a person-specific systematic error that does not average out over repeated cuffs. Recordings from wearables in daily life differ in several ways that are likely to increase the error: motion artifacts distort the pulse contour on which the second-derivative and morphology features depend; the sensor position (wrist instead of finger), contact pressure and skin perfusion change the PPG amplitude and shape independently of BP; and posture, the height of the sensor relative to the heart, temperature and physical activity change both the PPG and the BP. A device would therefore need a stricter signal-quality check, which reduces the number of usable segments, and a recalibration after changes in posture or sensor placement. None of these conditions is represented in our data, so the reported errors should be regarded as a lower bound for daily-life use. Second, the external database comprised only 17 cases, although each contributed long recordings with large pressure changes. Third, MIMIC-BP contains few segments above 180 mmHg (17 in the training pool, 3 in the test set), because its curation excluded systolic values above 200 mmHg [25]; severe hypertension could not be evaluated. Fourth, in MIMIC-BP, the features were computed from the first 10 s of each 30 s segment while the reference refers to the whole segment. Fifth, the original feature extractor also required an ECG channel; this did not exclude segments in our data (Section 2.1), but a PPG-only implementation would apply only the PPG checks and could retain slightly different segments. Sixth, the sampling rate of the recordings is higher than that of many wearables, and the effect of down-sampling on the features was not examined. Seventh, the latency was measured in Python on a desktop CPU and the memory was estimated; on-device timing and energy consumption remain to be measured, and the cost of pre-processing and feature extraction was estimated rather than measured on a device (Section 3.6). Confirming these estimates will require a fixed-point or integer implementation of the complete pipeline (filtering, beat delineation, feature computation and the gradient-boosted model) on a representative microcontroller (e.g., a Cortex-M4/M7-class device), with wall-clock latency, peak RAM and energy per estimate measured directly; we did not have access to such hardware for the present study, and this remains for follow-up work. Finally, the analysis is retrospective and the AAMI/ISO criteria are reported only for reference; a prospective validation following the standards for cuffless devices [22,23] is required before clinical use.

5. Conclusions

On unseen users, PPG features without calibration estimate blood pressure little better than a constant, regardless of the regressor. A delta model that predicts the change from a cuff reading is the most accurate calibration method, but its error grows with the time since the cuff, and periodic recalibration, not filtering or event-triggered cuffs, keeps it low; in the surgical data, recalibration nominally every 15 min gave about 10 mmHg SBP error. Because accuracy is limited by the information in the features rather than by model capacity,

a compact gradient-boosted model of about 47 KB was sufficient in our data, whereas large random forests and extrapolating linear or neural models were not. These results provide a quantitative basis for designing the calibration schedule and the model of PPG-based blood pressure monitoring on smart devices; confirming the recalibration interval in ambulatory wrist recordings and measuring on-device latency and energy are the next steps. The contribution of this study is a quantitative assessment of what calibration achieves and where it stops: the recency of the cuff, the recalibration interval and the cuff measurement error largely determine the accuracy. New-onset high-SBP segments were discriminated in MIMIC-BP, but no clinically usable alert was established; lower alert thresholds require independent validation, particularly given the poor external alarm performance. The SBP estimates did not meet the accuracy criteria for BP measuring devices in any configuration, and cuff measurement error further reduced the attainable accuracy. Errors were largest precisely at the high and low ends of the blood-pressure range where accuracy matters most (Section 3.3), and external validation rested on only 17 VitalDB cases; both the alarm-performance estimates and the external accuracy figures should therefore be regarded as preliminary until confirmed on a larger, independent cohort.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/diagnostics16223713/s1.

Author Contributions: Conceptualization, S.L.; methodology, S.L.; software, S.L.; validation, S.L.; formal analysis, S.L.; investigation, S.L.; data curation, S.L.; writing—original draft preparation, S.L.; writing—review and editing, S.L., M.A.A.-a. and G.L.; visualization, S.L.; supervision, G.L.; project administration, S.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Research Grant of Kwangwoon University in 2025.

Institutional Review Board Statement: This study was a secondary analysis of existing, de-identified data from MIMIC-BP and VitalDB. No new participant recruitment, contact, intervention or data collection was undertaken, and the authors had no access to identifiable information. The acquisition and release of VitalDB data were approved by the Institutional Review Board of Seoul National University Hospital (H-1408-101-605), with written informed consent waived because of data anonymity [27]. MIMIC-BP is derived from MIMIC-III, whose development was approved by the Institutional Review Boards of Beth Israel Deaconess Medical Center and the Massachusetts Institute of Technology. Individual patient consent was waived because the project did not affect clinical care and all protected health information was de-identified [26]. These approvals concern the source databases and do not constitute separate institutional approval of the present secondary analysis.

Informed Consent Statement: For the original VitalDB data collection, written informed consent was waived by the Seoul National University Hospital Institutional Review Board because of data anonymity. For MIMIC-III, from which MIMIC-BP was derived, individual patient consent was waived by the original institutional review boards because the project did not affect clinical care and protected health information was de-identified. The present secondary analysis involved no participant contact or collection of new identifiable data.

Data Availability Statement: MIMIC-BP [25] and VitalDB [27] are publicly available from their providers under their respective data use agreements; the authors are not permitted to redistribute them. The analysis code, including the feature extractor and the scripts that reproduce the subject splits, models and figures after downloading the data, is provided to the reviewers as Supplementary Material S1 and will be made available on GitHub upon publication and archived with a Zenodo DOI.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Feature List

Table A1. The 65 PPG features used in this study. The ECG-based features of the earlier extractor [28] (heart rate and three pulse arrival times) were not used.

Group	Features	*n*	Source
Previously published features (39)			
Temporal	ST, DT, CT; st_10, st_25, st_33, st_50, st_66, st_75; dt_10, dt_25, dt_33, dt_50, dt_66, dt_75	15	[28]
Amplitude/intensity	PIR, PPG_k, ppg_max, ppg_min	4	[28]
Derivative (VPG/APG)	dppgH, dppgW, ddppgH, ddppgW, ddppgPH, ddppgFH	6	[8,28]
Pulse-wave indices	AI, LASI, IPA	3	[28]
Area	S1, S2, S3, S4, As, Ad, areaPPG	7	[28]
Spectral	pxx_max (maximum of the multitaper PSD)	1	[28]
Statistical	ppg_skew, ppg_kurt, ppg_ent	3	[28]
Standard signal descriptors (10)			
Amplitude	PPG_pulse (peak minus foot), ps (peak), pd (foot), pm (mean over the beat)	4	this work
Spectral	max_power (maximum of the periodogram)	1	this work
Energy	TKE_PPG (mean Teager–Kaiser energy)	1	[29]
Hjorth	Hjorth_Act, Hjorth_Mob, Hjorth_Comp	3	[30]
Crossings	ZCR_PPG (zero-crossing rate of the mean-removed PPG)	1	this work
Morphology descriptors (16)			
Pulse contour	RI, dT	2	[31]
Pulse contour	notch_rel, pw50, up50, RI_sd	4	this work
Acceleration PPG	b/a, c/a, d/a, e/a, AGI	5	[8,32]
Gaussian decomposition	g_A21, g_dmu, g_s1, g_s2, g_r2	5	this work

Appendix B. Errors per VitalDB Case

Table A2. SBP errors per VitalDB case (delta model, mean over ten repeats; mmHg). *n*: evaluated segments; SD: within-case SD of the reference SBP; ME: mean error.

Case	*n*	Ref. SBP Mean (SD)	No Calibration MAE	1 Cuff MAE (ME)	Recal. 15 min MAE (ME)	Recal. 15 min DBP MAE
0001	775	129.2 (27.3)	22.8	18.0 (-1.0)	16.3 (−9.7)	7.4
0004	1686	108.2 (17.1)	17.4	12.4 (+1.7)	9.9 (−0.9)	4.1
0005	1880	113.6 (14.9)	12.2	23.5 (+22.9)	10.0 (+3.1)	5.8
0007	1271	122.6 (14.4)	10.7	9.8 (+1.7)	8.6 (−1.3)	5.8
0010	1724	130.7 (27.7)	25.5	26.5 (−21.5)	19.9 (−10.9)	5.9
0013	724	106.7 (11.4)	12.9	12.2 (+8.2)	9.6 (+3.3)	7.0
0016	1050	124.1 (16.3)	14.2	16.8 (−11.8)	10.5 (−3.7)	5.1
0017	1760	108.5 (15.1)	12.8	10.8 (−4.8)	9.9 (−0.0)	3.9
0019	2477	111.0 (13.9)	10.3	9.6 (+1.0)	8.0 (+1.0)	4.9
0020	2253	104.6 (10.5)	13.2	8.5 (+4.2)	6.5 (+3.7)	3.8
0022	1241	125.7 (19.2)	17.3	18.8 (−11.6)	9.8 (−5.3)	4.9
0024	384	134.0 (26.2)	26.3	31.8 (−29.3)	16.4 (−7.5)	8.0
0025	1268	114.2 (16.5)	13.3	15.0 (+8.8)	10.6 (+2.3)	6.1
0026	683	110.9 (12.1)	9.3	13.4 (+10.5)	8.9 (+4.9)	6.6
0027	1124	112.4 (15.9)	13.8	9.2 (−1.3)	10.2 (+1.0)	4.6
0028	2225	112.1 (14.0)	11.2	11.2 (−1.1)	10.2 (+1.1)	5.8
0029	1647	115.3 (12.9)	10.9	12.5 (+9.4)	8.6 (+0.5)	5.3

References

1. Mills, K.T.; Stefanescu, A.; He, J. The global epidemiology of hypertension. *Nat. Rev. Nephrol.* **2020**, *16*, 223–237. [CrossRef] [PubMed]
2. Mancia, G.; Kreutz, R.; Brunström, M.; Burnier, M.; Grassi, G.; Januszewicz, A.; Muiesan, M.L.; Tsioufis, K.; Agabiti-Rosei, E.; Algharably, E.A.; et al. 2023 ESH Guidelines for the management of arterial hypertension. *J. Hypertens.* **2023**, *41*, 1874–2071. [CrossRef] [PubMed]
3. Whelton, P.K.; Carey, R.M.; Aronow, W.S.; Casey, D.E.; Collins, K.J.; Dennison Himmelfarb, C.; DePalma, S.M.; Gidding, S.; Jamerson, K.A.; Jones, D.W.; 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. *J. Am. Coll. Cardiol.* **2018**, *71*, e127–e248. [CrossRef] [PubMed]
4. Poon, C.C.Y.; Zhang, Y.T. Cuff-less and noninvasive measurements of arterial blood pressure by pulse transit time. *Conf. Proc. IEEE Eng. Med. Biol. Soc.* **2005**, 5877–5880. [CrossRef] [PubMed]

5. Sharma, M.; Barbosa, K.; Ho, V.; Griggs, D.; Ghirmai, T.; Krishnan, S.K.; Hsiai, T.K.; Chiao, J.-C.; Cao, H. Cuff-Less and Continuous Blood Pressure Monitoring: A Methodological Review. *Technologies* **2017**, *5*, 21. [CrossRef]
6. Mukkamala, R.; Hahn, J.O.; Inan, O.T.; Mestha, L.K.; Kim, C.-S.; Töreyin, H.; Kyal, S. Toward ubiquitous blood pressure monitoring via pulse transit time: Theory and practice. *IEEE Trans. Biomed. Eng.* **2015**, *62*, 1879–1901. [CrossRef] [PubMed]
7. Kachuee, M.; Kiani, M.M.; Mohammadzade, H.; Shabany, M. Cuffless blood pressure estimation algorithms for continuous health-care monitoring. *IEEE Trans. Biomed. Eng.* **2017**, *64*, 859–869. [CrossRef] [PubMed]
8. Elgendi, M. On the analysis of photoplethysmogram signals. *Curr. Cardiol. Rev.* **2012**, *8*, 14–23. [CrossRef] [PubMed]
9. Qiu, Z.; Chen, D.; Ling, B.W.K.; Liu, Q.; Li, W. Joint regression network and window function-based piecewise neural network for cuffless continuous blood pressure estimation only using single photoplethesmogram. *IEEE Trans. Consum. Electron.* **2022**, *68*, 236–260. [CrossRef]
10. Liang, Y.; Chen, Z.; Ward, R.; Elgendi, M. Photoplethysmography and deep learning for blood pressure assessment: A review. *npj Digit. Med.* **2018**, *1*, 60.
11. Slapnicar, G.; Lustrek, M.; Gams, M. Blood pressure estimation from photoplethysmogram using a spectro-temporal deep neural network. *Sensors* **2019**, *19*, 3420. [CrossRef] [PubMed]
12. Mousavi, S.S.; Firouzmand, M.; Charmi, M.; Hemmati, M.; Moghadam, M.; Ghorbani, Y. Blood Pressure Estimation from Appropriate and Inappropriate PPG Signals Using a Whole-Based Method. *Biomed. Signal Process. Control* **2019**, *47*, 196–206. [CrossRef]
13. Chen, Q.; Yang, X.; Chen, Y.; Han, X.; Gong, Z.; Wang, D.; Zhang, J. A Blood Pressure Estimation Approach Based on Single-Channel Photoplethysmography Differential Features. *Biomed. Signal Process. Control* **2024**, *97*, 106662. [CrossRef]
14. Huang, B.; Chen, W.; Lin, C.L.; Juang, C.F.; Wang, J. MLP-BP: A Novel Framework for Cuffless Blood Pressure Measurement with PPG and ECG Signals Based on MLP-Mixer Neural Networks. *Biomed. Signal Process. Control* **2022**, *73*, 103404. [CrossRef]
15. Long, W.; Wang, X. BPnet: A Multi-Modal Fusion Neural Network for Blood Pressure Estimation Using ECG and PPG. *Biomed. Signal Process. Control* **2023**, *86*, 105287. [CrossRef]
16. Huang, B.; Wei, Y.; Ren, J.; Cheng, Y.; Zheng, Z. Continuous Blood Pressure Monitoring Based on Transformer Encoders and Stacked Attention Gated Recurrent Units. *Biomed. Signal Process. Control* **2025**, *99*, 106860. [CrossRef]
17. Tian, Z.; Liu, A.; Zhu, G.; Chen, X. A Paralleled CNN and Transformer Network for PPG-Based Cuff-Less Blood Pressure Estimation. *Biomed. Signal Process. Control* **2025**, *99*, 106741. [CrossRef]
18. González, S.; Hsieh, W.T.; Chen, T.P.C. A benchmark for machine-learning based non-invasive blood pressure estimation using photoplethysmogram. *Sci. Data* **2023**, *10*, 149. [CrossRef] [PubMed]
19. Payne, R.A.; Symeonides, C.N.; Webb, D.J.; Maxwell, S.R.J. Pulse transit time measured from the ECG: An inadequate surrogate for blood pressure. *J. Hum. Hypertens.* **2006**, *20*, 45–54.
20. Zhang, G.; Gao, M.; Xu, L.; Mukkamala, R. Pulse arrival time is not an adequate surrogate for pulse transit time as a marker of blood pressure. *J. Appl. Physiol.* **2011**, *111*, 1681–1686. [CrossRef] [PubMed]
21. Cohen, J.B.; Byfield, R.L.; Hardy, S.T.; Juraschek, S.P.; Miller, N.H.; Mukkamala, R.; Picone, D.S.; Thiele, R.H.; Yang, E.; Brady, T.M. Cuffless devices for the measurement of blood pressure: A scientific statement from the American Heart Association. *Hypertension* **2026**, *83*, 876–886. [CrossRef] [PubMed]
22. *IEEE 1708-2014*; IEEE Standard for Wearable, Cuffless Blood Pressure Measuring Devices. IEEE: New York, NY, USA, 2014.
23. *ISO 81060-3:2022*; Non-Invasive Sphygmomanometers—Part 3: Clinical Investigation of Continuous Automated Measurement Type. ISO: Geneva, Switzerland, 2022.
24. Wu, Y.; Li, Z.; Jiang, C.; Yu, H.; Mao, Y.; Liu, K.; Yao, J. Reliability-Aware fusion of acoustic signals and electrical impedance tomography for respiratory functional assessment. *Measurement* **2026**, *290*, 123020. [CrossRef]
25. Sanches, I.; Gomes, V.V.; Caetano, C.; Cabrera, L.S.B.; Cene, V.H.; Beltrame, T.; Lee, W.; Baek, S.; Penatti, O.A.B. MIMIC-BP: A curated dataset for blood pressure estimation. *Sci. Data* **2024**, *11*, 1233. [CrossRef] [PubMed]
26. Johnson, A.E.W.; Pollard, T.J.; Shen, L.; Lehman, L.H.; Feng, M.; Ghassemi, M.; Moody, B.; Szolovits, P.; Celi, L.A.; Mark, R.G. MIMIC-III, a freely accessible critical care database. *Sci. Data* **2016**, *3*, 160035. [CrossRef] [PubMed]
27. Lee, H.C.; Park, Y.; Yoon, S.B.; Yang, S.M.; Park, D.; Jung, C.-W. VitalDB, a high-fidelity multi-parameter vital signs database in surgical patients. *Sci. Data* **2022**, *9*, 279. [CrossRef] [PubMed]
28. Lee, S.; Al-Antari, M.A.; Joshi, G.P. Improved Confidence-Interval Estimations Using Uncertainty Measure and Weighted Feature Decisions for Cuff-Less Blood-Pressure Measurements. *Bioengineering* **2025**, *12*, 131. [CrossRef] [PubMed]
29. Kaiser, J.F. On a simple algorithm to calculate the 'energy' of a signal. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Albuquerque, NM, USA, 3–6 April 1990; pp. 381–384.
30. Hjorth, B. EEG analysis based on time domain properties. *Electroencephalogr. Clin. Neurophysiol.* **1970**, *29*, 306–310. [CrossRef] [PubMed]
31. Finnegan, E.; Davidson, S.; Harford, M.; Watkinson, P.; Tarassenko, L.; Villarroel, M. Features from the photoplethysmogram and the electrocardiogram for estimating changes in blood pressure. *Sci. Rep.* **2023**, *13*, 986. [CrossRef] [PubMed]

32. Goda, M.Á.; Charlton, P.H.; Behar, J.A. pyPPG: A Python toolbox for comprehensive photoplethysmography signal analysis. *Physiol. Meas.* **2024**, *45*, 045001. [CrossRef] [PubMed]
33. Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794. [CrossRef]
34. O'Brien, E.; Petrie, J.; Littler, W.; De Swiet, M.; Padfield, P.L.; Altman, D.; Bland, M.; Coats, A.; Atkins, N. The British Hypertension Society protocol for the evaluation of blood pressure measuring devices. *J. Hypertens.* **1993**, *11*, S43–S62. [CrossRef] [PubMed]
35. Stergiou, G.S.; Alpert, B.; Dey, S.; Asmar, R.; O'Brien, E.; Shennan, A. AAMI/ESH/ISO Universal Standard for the validation of blood pressure measuring devices. *Hypertension* **2018**, *71*, 368–374. [CrossRef] [PubMed]
36. *ISO 81060-2:2018*; Non-Invasive Sphygmomanometers—Part 2: Clinical Investigation of Intermittent Automated Measurement Type. ISO: Geneva, Switzerland, 2018.
37. Pavan, M.; Pezzarossa, L.; Pittorino, F.; Roveri, M.; Fafoutis, X. What changes after deployment? A survey on on-device learning in TinyML. *arXiv* **2026**, arXiv:2605.31226. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.