scientific reports

Check for updates

OPEN

Novel metrics for tracking blood pressure changes incontinuous cuffless blood pressure estimations

Shan He✉ & Miodrag Bolić

Recent studies revealed the importance of tracking continuous blood pressure (BP) changes in monitoring and controlling hypertension and diagnosing cardiovascular diseases. However, current evaluation protocols utilize distance measures as primary metrics, which cannot properly evaluate the ability of the estimation model to track BP changes. This paper proposes a comprehensive evaluation framework which evaluates the distance and trend similarity metrics, and the composite metric of both between the reference and estimated BPs. The results of applying both widely used conventional metrics and the new proposed metrics for BP estimations are demonstrated in an example of comparing the reference with a set of different BP estimations. Then, the metrics are applied to BP estimations obtained using state-of-the-art (SOTA) algorithms. It is shown that even though SOTA algorithms have a low mean and standard deviation of absolute difference, they are not capable of tracking short-term blood pressure changes. Additionally, the proposed metrics are normalized metrics and range from -1 to 1, making them intuitively interpretable, similar to well-known correlation coefficients. Therefore, we suggest that the proposed evaluation framework should be used regularly in evaluating continuous BP monitoring systems.

Blood pressure (BP) is one of the vital signs-together with respiratory rate, heart rate, oxygen saturation, and body temperature-that can evaluate an individual's health[1]. Hypertension, which is the most common BP disorder, has been reported as the world's leading cause of death[2]. Continuous cuffless BP measurements, which estimate BP without a cuff and incorporate machine learning, deep learning, and wearable sensors, have become a popular research problem in the past decade. The promise of this new technique is that it addresses the pain points of conventional cuff-based devices (e.g., auscultatory[3], oscillometric[4]) in monitoring long-term BP continuously and comfortably. Therefore, besides intermittent systolic and diastolic BP (SBP and DBP) measurements, cuffless BP estimation is expected to be applied for: (1) continuous and long-term BP monitoring, tracking BP patterns and BP variability (BPV) for early diagnosis of BP disorders (e.g., hypertension and hypotension) and cardiovascular diseases (CVD); (2) nocturnal BP measurement which is hard to measure by conventional devices; (3) long-term BP management follow-up (e.g., the effect of anti-hypertension medication); (4) incorporating with other wearable sensors (e.g., body temperature, acceleration, respiration) for detailed hemodynamic and overall health assessment.

There have been several different technologies applied for cuffless BP estimation[5]. For example, a non-contact BP estimation method was developed by processing facial video and extracting pulse waveform features using a smartphone[6]. The ultrasound sensor was applied to measure the cross-sectional area and blood velocity of the carotid artery and compute pulse pressure based on fluid dynamic principles[7,8]. Recently, the most popular method involves employing a pulse wave analysis (PWA) approach, either with or without pulse transit time (PTT), integrated with deep learning models. The work[9] proposed a non-linear model to predict BP using pulse arrival time (PAT) which is considered as a surrogate to PTT and it achieved the root mean squared difference (RMSD) of 3.43 mmHg and 3.32 mmHg for SBP and DBP, respectively. Hsu *et al*[10] extracted 59 features from the PPG signal and its first derivative and selected 32 features as the input of a fully-connected deep neural network for BP estimation. The RMSD for SBP and DBP is 4.63 mmHg and 3.0 mmHg, respectively. The morphology of the PPG and BP waveforms has also been proven highly correlated[11], which indicates the feasibility of modeling BP waveforms directly from a simultaneous PPG waveform. A U-Net-based deep learning model was developed to approximate BP waveforms from PPG waveforms and achieved satisfactory results, with MAD±SDAD (mean and standard deviation of absolute difference) of (5.73 ± 9.16) mmHg and (3.45 ± 6.15) mmHg for SBP and DBP respectively[12].

Health Devices Research Group, School of Electrical Engineering and Computer Science, Faculty of Engineering, University of Ottawa, Ottawa K1N 6N5, Canada. ✉email: she014@uottawa.ca

	ANSI/AAMI SP10:2002[14]	ISO Standard81060-2:2018[16]	IEEE Standard1708a-2019[20]	ISO Standard81060-3:2022[21]	ESHrecommendation 2023[22]
Device type	Cuff-based; intermittent	Cuff-based; intermittent	Cuffless; intermittent	Cuffless; continuous	Cuffless; continuous
Reference method	Ausculatory; intra-arterial	Mercury sphygmomanometers; accurate non-mercury	Ausculatory	Intra-arterial	Ausculatory24-h oscillometric
Main accuracy metrics	MD;SDD	MD; SDD	MAD	MD;SDD	MD;SDD
Pass requirement (mmHg)	$-5 \leq \mathrm{MD} \leq 5$ SD ≤ 8	$-5 \leq \mathrm{MD} \leq 5$ SD ≤ 8	MAD ≤ 7	$-5 \leq \mathrm{MD} \leq 5$ SD ≤ 8	$-5 \leq \mathrm{MD} \leq 5$ SD ≤ 8

Table 1. Key aspects of validation protocols developed for cuff-based and cuffless BP devices. *MD* mean estimation difference, *SDD* standard deviation of estimation difference. See Table 2, *MAD* mean absolute estimation difference. See Table 2.

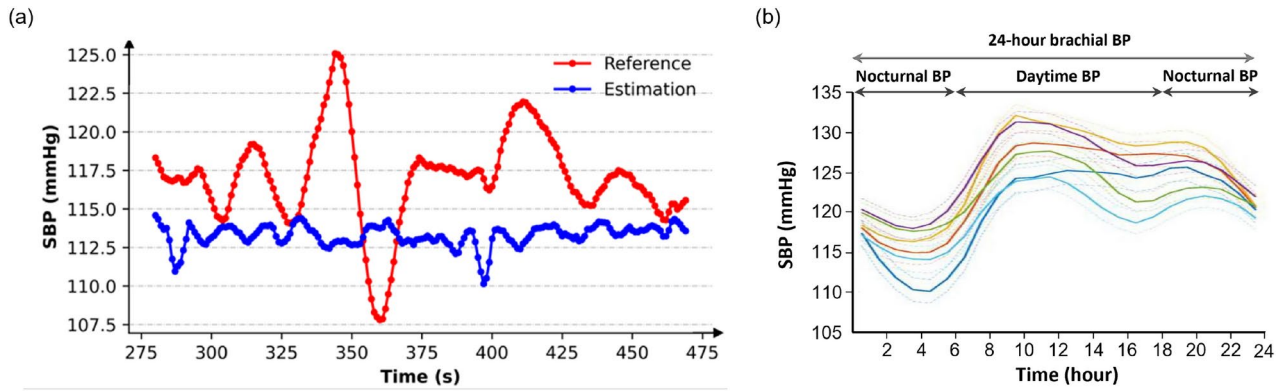


Fig. 1. (**a**) An example of the proposed BP model unable to track BP changes but the overall MAD and SDAD is low. This segment is 543 s, MAD±SDAD for SBP is (3.77 ± 2.51) mmHg. (**b**) 24-h ambulatory brachical SBP pattern, stratified by age. Solid lines are mean values and dash lines are 95% confidence intervals (reproduction of the middle figure of Fig. 2 demonstrated in work[23], author authorized the use of this figure in this study).

Although the current cuffless BP studies have shown promising results, existing performance validation standards are inadequate for this new research problem. In the last three decades, several organizations, such as the British Hypertension Society (BHS)[13], Association for Advancement of Medical Instrumentation (AAMI)[14], and the European Society of Hypertension (ESH)[15] have proposed their individual protocols of clinical validation for BP measurement devices. In 2018, a single universal standard (AAMI/ESH/ISO Universal Standard - ISO 81060-2:2018[16]) was developed by AAMI, ESH and ISO experts for global use of BP device validation and it was expected to replace all other previous standards/protocols. Although these validation protocols were designed for cuff-based devices and for intermittent assessments, the majority of current cuffless BP studies[10,12,17] still employed these protocols to evaluate the performance of their proposed approaches, which was considered inadequate and misleading for continuous and cuffless BP measurement[18]. Therefore, new standards were adjusted and adapted based on aforementioned cuff-based protocols for cuffless BP devices specifically. In 2014, the IEEE (Institute of Electrical and Electronics Engineers) published the first standard for validating the accuracy of cuffless wearable BP devices (IEEE 1708-2014)[19] and an amendment of this protocol was updated in 2019 (IEEE 1708a-2019)[20]. ISO (International Organization for Standardization) published their protocol ISO 81060-3:2022[21] which is specific for continuous noninvasive sphygmomanometers in 2022. In 2023, EHS also proposed their recommendations for the validation of cuffless BP measuring devices[22]. A summary of key aspects of validation protocols from several aforementioned standards are listed in Table 1.

In Table 1, several similarities can be observed among different protocols for both cuff-based and cuffless devices. The pairwise distances between the estimations and references are used as the primary metrics to validate the performance of the proposed device against the reference measurements. However, we observed from experiments (see Section S1 for details) that sometimes an estimation model may meet the distance criteria, even though it may not accurately monitor BP changes. For example, the MAD±SDAD of a 543 seconds' testing subset shown in Fig. 1a is (3.77 ± 2.51) mmHg for SBP which is considered as "Grade A" of IEEE Standard 1708a-2019[20]. However, the model is unable to track one acute BP change occurred between 325 and 375 seconds. We believe this is a general but underestimated problem of the existing studies. It is because the existing validation standards do not well define the metrics to evaluate the ability of the proposed methods/devices in tracking BP changes.

Recently, the increasing number of studies revealed the BP pattern is related to different diseases and it is also important to accurately track BP patterns for both short-term and long-term. BP pattern refers to the morphology of sequential BP readings from a certain period of time which consists a number of monotonous trends. For example, several studies[23–25] observed that BP has a daily pattern and a global experiment[23] of 144,509

brachial SBP measurements from 2423 healthy subjects is shown in Fig. 1b. It can be observed that BP starts to rise a few hours before a person wakes up and continuous to increase during the day, peaking in midday. Then BP typically decreases in the late afternoon and evening, and even lower at night while sleeping. An accurate measure of BP pattern can assess BP changes and BP variability. Based on the time window BP is assessed, BP pattern can be categorized and measured in different ranges, including very short-term (beat-to-beat, minute-to-minute), short-term (within 24 h, hour-to-hour, day-to-night), long-term (day-to-day, months and years)[26,27]. Abnormal BP patterns may indicate specific forms of hypertension (e.g. nocturnal hypertension) or may be associated with concomitant illness (e.g. sleep apnea) and recognition of these abnormal BP patterns can be helpful in diagnosis and management of hypertension[28]. Cardiovascular analysis indicates that obstructive sleep apnea (OSA) will cause acute hemodynamic changes[29]; both BP and heart rate (HR) initially decrease and then progressively increase by approximately 20 mmHg and 15 bpm, respectively[30]. Orthostatic hypotension, which indicates impaired physical performance and cerebral autoregulation, also causes rapid BP changes, SBP declines at least 20 mmHg and/or DBP declines at least 10 mmHg within 3 minutes (most declines occur within the first minute[31]) after standing up[32]. An example of BP change while a person stands up from sitting is demonstrated in Figure S1(d), where SBP rapidly decreases from 124 mmHg to 85 mmHg within 10 seconds and then recovers to normal. The work[33] indicates 24-hour BP values are linearly related to stroke risk and the Dublin Outcome study concluded that nocturnal BP was overall the best predictor of cardiovascular risk, with a 10 mmHg increase in mean nocturnal SBP being associated with a 21% increase in cardiovascular mortality[34]. It was also studied that sleep phases, such as rapid eye movement and non-rapid eye movement, have different BP patterns[35]. Therefore, it is vital for future studies to evaluate the performance of BP pattern tracking with their estimation models, and eventually facilitating the adoption of continuous cuffless BP estimation in real-world clinical applications.

The main contributions of this paper are summarized as follows:

1. To the best of our knowledge, this is the first study that reveals the limitations of current evaluation protocols in BP pattern tracking and comprehensively compares the performance of several widely used metrics in BP pattern tracking;
2. This study proposes a comprehensive evaluation framework for cuffless BP estimation which compares the similarities between the references and estimations in terms of both **distance** and **trend**. A composite metric is also proposed to combine both distance and trend similarity metrics. The proposed similarity metrics are normalized metrics which range from -1 to 1 and can be easily interpreted. The new proposed similarity metrics are generic and can be applied to other healthcare monitoring studies and even general time series similarity comparison problems.The rest of this paper is organized as follows. Several widely applied numerical metrics for evaluating distance similarity and shape similarity are introduced and compared. A comprehensive evaluation framework is proposed which compares the similarity between the references and estimations in terms of distance and trend, and a composite of both. The new proposed similarity metrics are applied to a set of simulated data to demonstrate their performance and then applied to the estimates obtained using SOTA algorithms.

Conventional performance metrics for cuffless BP evaluation

Distance metrics

The pairwise distance measures between the estimated and reference BPs are the most common metrics employed by the majority of current studies, as well as the aforementioned validation standards. For SBP and DBP, the mean and standard deviation of estimation difference (MD and SDD), the mean and standard deviation of absolute difference (MAD and SDAD) and the root mean squared difference (RMSD) of the *n* individual paired measurements of the estimation and of the reference measurements for all subjects are calculated using equations

Distance metrics			**Correlation coefficients**	
Estimation difference	**Mean**	**Standard deviation**	**Pearson correlation**	$\rho = \frac{\sigma_{12}}{\sigma_1\sigma_2}$
	$\mathrm{MD} = \frac{1}{n}\sum_{i=1}^{n}(\mathrm{BP}_{\mathrm{est}i} - \mathrm{BP}_{\mathrm{ref}i})$	$\mathrm{SDD} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\mathrm{BP}_{\mathrm{est}i} - \mathrm{BP}_{\mathrm{ref}i} - \mathrm{MD})^2}$		
Absolute difference	Mean	Standard deviation		
	$\mathrm{MAD} = \frac{1}{n}\sum_{i=1}^{n}\lvert\mathrm{BP}_{\mathrm{est}i} - \mathrm{BP}_{\mathrm{ref}i}\rvert$	$\mathrm{SDAD} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\lvert\mathrm{BP}_{\mathrm{est}i} - \mathrm{BP}_{\mathrm{ref}i}\rvert - \mathrm{MAD})^2}$	Concordance correlation	$\rho_c = \frac{2\sigma_{12}}{(\mu_1-\mu_2)^2+\sigma_1^2+\sigma_2^2}$
Root mean squared difference	$\mathrm{RMSD} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\mathrm{BP}_{\mathrm{est}i} - \mathrm{BP}_{\mathrm{ref}i})^2}$			

Table 2. Definition of distance metrics and correlation coefficients widely-used for cuffless BP estimation evaluation. $\mathrm{BP}_{\mathrm{est}}$: BP estimated by the test device/method; $\mathrm{BP}_{\mathrm{ref}}$: BP measured by the reference device/method. *i*: index of the individual element; *n*: the number of measurements σ_{12}: covariance of $\mathrm{BP}_{\mathrm{est}}$ and $\mathrm{BP}_{\mathrm{ref}}$; σ_1 and σ_2: standard deviation of $\mathrm{BP}_{\mathrm{est}}$ and $\mathrm{BP}_{\mathrm{ref}}$; μ_1 and μ_2: mean of $\mathrm{BP}_{\mathrm{est}}$ and $\mathrm{BP}_{\mathrm{ref}}$

shown in Table 2. Although these pairwise Euclidean distance based metrics are straightforward, intuitive and less computational complex, they can not properly evaluate the accuracy of BP pattern tracking. For instance, the values of these aforementioned distance metrics between a set of simulated BP estimations and the reference (see Fig. 6 and "Applying the proposed new metrics to simulated data" for details) are illustrated in Table 7. It can be observed that simulated BP estimations $\mathrm{BP}_{\mathrm{est1-4}}$ satisfy the passing requirement of IEEE Standard 1708a-2019 (MAD$\leq$7 mmHg), but $\mathrm{BP}_{\mathrm{est2-4}}$ were unable to accurately track BP patterns. Though MAD between $\mathrm{BP}_{\mathrm{est5}}$ and $\mathrm{BP}_{\mathrm{ref}}$ is 8.14 mmHg, but $\mathrm{BP}_{\mathrm{est5}}$ almost has the identical morphology to $\mathrm{BP}_{\mathrm{ref}}$ which performs better than $\mathrm{BP}_{\mathrm{est2-4}}$ in BP pattern tracking. It can also be observed that MAD and RMSD can better indicate the accuracy of the estimation model in BP change tracking than MD, for example, comparing $\mathrm{BP}_{\mathrm{est2}}$ and $\mathrm{BP}_{\mathrm{est3}}$ against $\mathrm{BP}_{\mathrm{ref}}$. Furthermore, the contributions of individual error values to the final results are different for MAD and RMSD, since the errors are squared before being averaged, the RMSD assigns a higher weight to large errors than MAD. For example, the RMSD between $\mathrm{BP}_{\mathrm{est4}}$ and $\mathrm{BP}_{\mathrm{ref}}$ shown in Fig. 6 is 6.24 mmHg whereas the MAD is 4.29 mmHg.

There are also some distance metrics developed to measure curve similarity, such as Dynamic time warping (DTW)[36], Fréchet distance[37] and Hausdoff distance[38]. These metrics address the key limitations of pairwise distance based metrics in terms of time scaling and shifting and time series in different lengths, but they are more computational complex than pairwise distance metrics. As the length of estimation and reference are identical and the time delay (may be caused by time series estimation models, such as RNN and LSTM) should be composited before comparison. These metrics are not necessary for cuffless BP evaluation.

Therefore, we suggest to use MAD or RMSD as the primary distance metric for future studies.

Correlation coefficient

A popular metric to evaluate the agreement between estimations and reference measurements is the correlation coefficient. Pearson correlation coefficient (PCC, denoted as ρ), which measures the linear correlation between two variables, is commonly employed by the majority of existing studies[10,39]. The mathematical definition of PPC is shown in Table 2. However, ρ has one major shortcoming when assessing the reproducibility of measurements: it is invariant to additive or/and multiplicative shifts by a constant value, which is also an issue for the Spearman correlation coefficient. Therefore, using ρ to compare the similarity of the estimated BPs and the reference BPs may be misleading.

The concordance correlation coefficient (CCC, denoted as ρ_c) was originally introduced by Lawrence Lin to address the aforementioned limitations of PCC in 1989[40], and the formulation of CCC is shown in Table 2. The relationship between ρ and ρ_c can be further derived as $\rho_c = \rho \cdot C_{\mathrm{b}}$ (see Section S3 for details) which means CCC (ρ_c) contains the measurements of both C_{b} (a measure of accuracy) and PCC (a measure of precision). A demonstration of limitations of PCC and the robustness of CCC in detecting scale and location shifts comparing to PCC is shown in Section S3.

The previous works[41,42] translate ρ into descriptors like 'weak," "moderate," or "strong", and McBride[43] suggested that $\rho_c > 0.99$ for "almost perfect," $0.95 < \rho_c \leq 0.99$ for "substantial," $0.90 < \rho_c \leq 0.95$ for "moderate" and $\rho_c \leq 0.90$ for "poor." However, there is still no clear-cut agreement how to interpret both PCC and CCC for cuffless BP evaluation.

Though the results of PCC and CCC shown in Table 7 can indicate the morphology similarities of a simulated dataset. They are not suitable for evaluating BP pattern tracking of noisy and long-term data collected from real subjects, as described in "Combining the segment similarity measures (global evaluation)" and "Applying the proposed new metrics to results of two SOTA models".

A composite score to represent similarity in mean, variance and correlation coefficient

A composite score, generic composite similarity measure (GCSM)[44], which was originally proposed to compare similarities of spatial or temporal patterns from images for ecological studies can be used to address the limitation that a single distance metric can not measure the pattern similarity for two compared sequences (as described in "Distance metrics"). GCSM composes the similarity measures of mean, standard deviation and correlation coefficient of two compared sequences (x and y) to a single score, and its formulation is shown in Table 3.

It can be inferred from the formulations that $s_1(x,y) \in [0,1]$, $s_2(x,y) \in [0,1]$ and $s_3(x,y) \in [-1,1]$ and therefore $\mathrm{GCSM}(x,y) \in [-1,1]$. The mean, standard deviation and PCC are more similar for two compared sequences when $\mathrm{GCSM}(x,y)$ runs to 1. Although the GCSM consists of three different metrics to measure

$\mathbf{GCSM}(x,y) = s_1(x,y) \cdot s_2(x,y) \cdot s_3(x,y)$		
Components		
$s_1(x,y) = \begin{cases} 1 - \frac{\lvert\mu_x - \mu_y\rvert}{\max_{x,y} - \min_{x,y}} & s_3 \geq 0 \\ 1 - \frac{\lvert\min_{x,y} + \max_{x,y} - \mu_x - \mu_y\rvert}{\max_{x,y} - \min_{x,y}} & s_3 < 0 \end{cases}$	$s_2(x,y) = 1 - \frac{\lvert\sigma_x - \sigma_y\rvert}{(\max_{x,y} - \min_{x,y})/2}$	$s_3(x,y) = \begin{cases} 1 & \sigma_x = \sigma_y = 0 \\ 0 & \sigma_x \text{ or } \sigma_y = 0 \\ \frac{\sigma_{x,y}}{\sigma_x \sigma_y} & \text{otherwise} \end{cases}$

Table 3. Equation for calculating GCSM and definitions of each component. μ_x and μ_y: mean values of the compared sequences x and y; $\max_{x,y}$ and $\min_{x,y}$ are the global maximum and minimum of sequences x and y; σ_x and σ_y: standard deviation of two sequences x and y; σ_{xy} is the covariance of sequences x and y.

the overall similarity between sequences, it still has drawbacks. $s_1(x, y)$ and $s_2(x, y)$ are sensitive to outliers if $|\mu_x - \mu_y| \ll \max_{x,y} - \min_{x,y}$ or $|\sigma_x - \sigma_y| \ll (\max_{x,y} - \min_{x,y})/2$, the $s_1(x, y)$ or $s_2(x, y)$ can still be close to 1 which indicates two sequences have similar means and standard deviations, though the actual differences (the numerators of $s_1(x, y)$ and $s_2(x, y)$) can be significantly large and cannot be ignored.

A new comprehensive performance evaluation for cuffless BP estimation

In this section, a new comprehensive performance evaluation method which quantifies the similarity of **distance**, **trend** and the composite of them, between the reference and estimated BP measurements will be introduced. This new method addresses the problems or limitations of the existing evaluation metrics and methods employed for cuffless BP estimation (as described in "Conventional performance metrics for cuffless BP evaluation"). The workflow of this new method is demonstrated in Fig. 2, and the pseudocode of the proposed method is shown in Algorithm S1. The pre-processing is done with piecewise linear representation (PLR) which divides the compared estimated and reference BP sequences into segments according to trend changes and approximates segment linear trends. For each segment, new proposed distance and trend similarity metrics will be evaluated, and a composite similarity metric which combines both distance and trend similarities using weighted sum is then calculated. Finally, the temporal normalization method will combine the segment similarity results for entire sequence evaluation and result interpretation.

Pre-processing

The similarity of two raw time series is difficult to measure as time series data are typically noisy and of high dimensions. As a result, it is common and efficient to work with data representations that are lighter than the raw data by reducing the dimensions of the data while maintaining their main properties, a number of techniques can be applied for this purpose (see study[45] for methods introduction and comparison). In this study, piecewise linear representation (PLR)[46] which can effectively extract trend information from the original time series data by dividing the original data into a number of segments according to trend changes and approximating the linear trends is employed. The number of segments that one time series should be divided into can be decided using several approaches, such as sliding window[47], top-down[48] and bottom-up[49] algorithms (see[50] for details). However, these methods are parametric which means parameters such as the number of segments to be divided into or the approximation error threshold must be pre-defined, and these parameters will impact the segmentation results and therefore affect similarity comparison between sequences (e.g., the results demonstrated in Table 2 of study[51]). In this study, a non-parametric Bayesian method, Bayesian estimator of abrupt change, seasonality, and trend (BEAST)[52], is employed to pre-process the original compared sequences. A demonstration of detecting change points and segmenting the reference and estimated BPs using BEAST is shown in Fig. 5a. However, two time series may have different number of segments (e.g., the reference BP is divided into 8 segments and estimated BP is divided into 9 segments as shown in Fig. 5a), then a equal segment number (ESN) process which re-divides both time series into the same number of segments while also ensures a segment has enough samples ($m \geq 2$) for linear trend approximation is developed and applied to the compared sequences.

After segmentation and ESN process, the least squares method is applied to approximate the original segment data to a straight line which represents the segment trend (pre-processing steps are described in Algorithm S1 line1 to line4). A demonstration of ESN process and linear trend approximation is shown in Fig. 5b.

To better describe the proposed similarity metrics in the following sections, a set of denotations are formally introduced. The original reference sequence (reference BPs) is denoted as x and the original estimated sequence (estimated BPs) is denoted as y respectively. The i-th segment of the original sequences x and y are denoted as x_i and y_i respectively. The approximated linear representations of the original sequences x and y using aforementioned PLR methods are denoted as x_L and y_L respectively. Finally, the i-th segment of the linear trends of the compared sequences are denoted as x_{L_i} and y_{L_i} respectively.

Segment similarity measures

After pre-processing two compared sequences using PLR, the segment similarity measures of distance, trend and the composite of them will be evaluated.

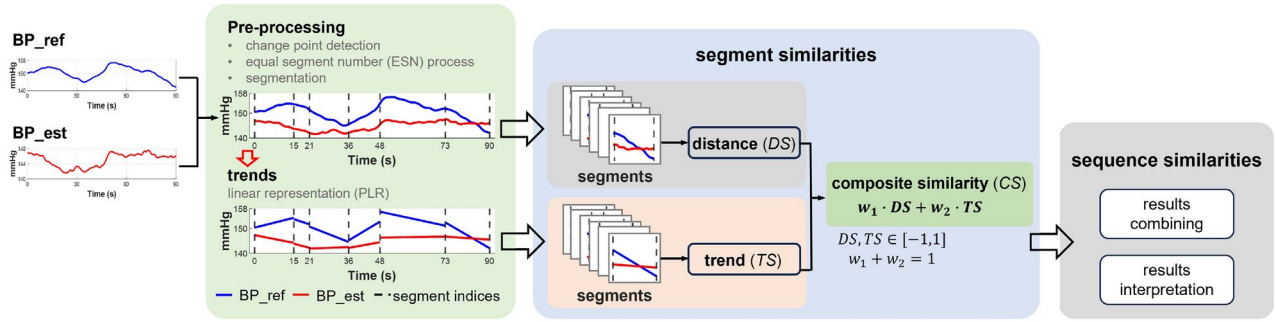


Fig. 2. The framework of the proposed comprehensive performance evaluation method for cuffless BP estimation, including sequence pre-processing, segment similarity evaluation and sequence similarity evaluation.

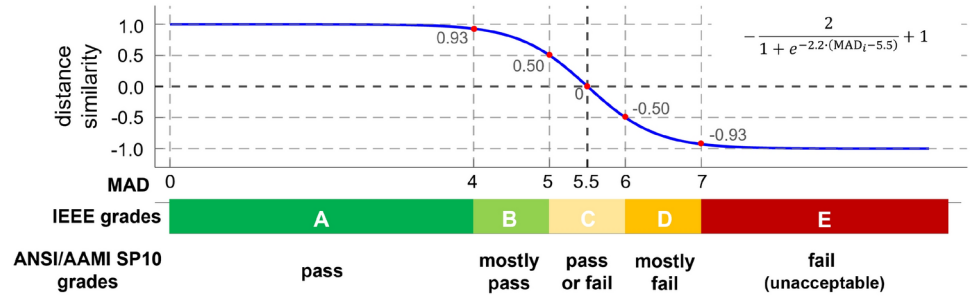


Fig. 3. A demonstration of the proposed segment distance similarity which is a customized logistic function according to IEEE Standard 1708a-2019 evaluation criteria[20].

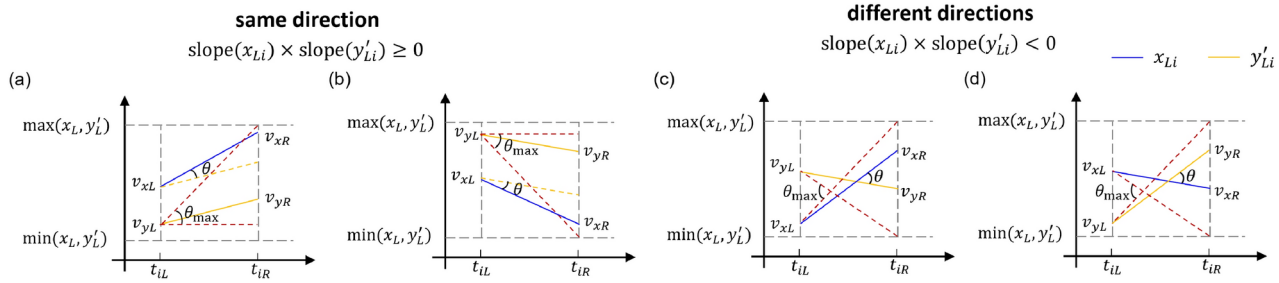


Fig. 4. A demonstration of the proposed angular difference based trend similarity measures of approximated linear segments in the same direction and in opposite directions, with the corresponding calculations of θ and $\theta_{\max}$.

Segment distance similarity

The new proposed segment distance similarity evaluates the MAD between the estimated and reference BPs, and maps the MAD to a value which is between -1 to 1. The distance similarity metric is a logistic function customized according to IEEE Standard 1708a-2019[20] evaluation protocol as shown in Fig. 3. For each segment, the MAD between two compared sequences is calculated and then the segment distance similarity $DS(x_i, y_i)$ is defined as:

$$DS(x_i, y_i) = -\frac{2}{1 + e^{-2.2\cdot(\mathrm{MAD}_i - 5.5)}} + 1 \tag{1}$$

where MAD_i is the MAD of the i-th segment of compared sequences after change point detection and segmentation, $\mathrm{MAD}_i = \frac{\sum_{j=1}^{m} |x_{ij} - y_{ij}|}{m}$, x_{ij} and y_{ij} are the j-th measurement values of the i-th segment of compared sequences x and y respectively, and m is number of samples of the i-th segment ($m \geq 2$). It can be inferred from Eq. (1) and observed from Fig. 3 that $DS(x_i, y_i) \in [-1, 1]$. $DS(x_i, y_i)$ is close to 1 when $\mathrm{MAD}_i \leq 4$ mmHg, corresponding to "Grade A" of IEEE Standard and "pass" grade of ANSI/AAMI SP10[14] requirement. $DS(x_i, y_i) \in [0.5, 0.93)$ when $4\mathrm{mmHg} < \mathrm{MAD}_i \leq 5\mathrm{mmHg}$, corresponding to "Grade B" of IEEE Standard and "mostly pass" grade of ANSI/AAMI SP10 standard respectively. $DS(x_i, y_i) \in [-0.5, 0.5)$ when $5\mathrm{mmHg} < \mathrm{MAD}_i \leq 6\mathrm{mmHg}$ ($DS(x_i, y_i) = 0$ when $\mathrm{MAD}_i = 5.5$ mmHg), corresponding to "Grade C" of IEEE Standard and "pass or fail" grade of ANSI/AAMI SP10 standard respectively. $DS(x_i, y_i) \in [-0.93, -0.5)$ when $6\mathrm{mmHg} < \mathrm{MAD}_i \leq 7\mathrm{mmHg}$, corresponding to "Grade D" of IEEE Standard and "mostly fail" grade of ANSI/AAMI SP10 standard respectively. $DS(x_i, y_i)$ is close to -1 when $\mathrm{MAD}_i \geq 7\mathrm{mmHg}$, corresponding to "Grade D (unacceptable)" of IEEE Standard and "fail" grade of ANSI/AAMI SP10 standard respectively.

Segment trend similarity

The new proposed segment trend similarity metric evaluates the accuracy of the estimation method in tracking BP trends. Firstly, the approximated linear representation of the estimated BP y_L is processed such that it has the same mean to the approximated linear representation of the reference BP x_L ($y'_L = y_L - (\overline{y_L} - \overline{x_L})$, $\overline{x_L}$ and $\overline{y_L}$ are the mean values of the approximated linear representations for reference BP x_L and estimated BP y_L respectively). For each segment, the slopes of the approximated linear trends are calculated to determine if the compared segments are in the same direction or not. If two compared segment trends are in the same direction, as showing in Fig. 4a, b, both of the segments are up or down, or at least one segment is unchanged ($\mathrm{slope}(x_{L_i}) \times \mathrm{slope}(y'_{L_i}) \geq 0$). Therefore, the range of angular difference θ between the segment linear trends

x_{L_i} and y'_{L_i} is $[0, \theta_{\max}]$ and the segment trend similarity is defined as $TS(x_i, y_i) = 1 - \frac{\theta}{\theta_{\max}}$. If two compared segment trends are in different directions, as showing in Fig. 4c, d, one segment is up and the other segment is down ($\text{slope}(x_{L_i}) \times \text{slope}(y'_{L_i}) < 0$). Therefore, the range of angular difference θ between segment linear trends x_{L_i} and y'_{L_i} is $(0, \theta_{\max}]$ and the segment trend similarity is defined as $TS(x_i, y_i) = -\frac{\theta}{\theta_{\max}}$. The formulations of θ and $\theta_{\max}$ for different scenarios are illustrated in Table 4 and Fig. 4.

As shown in Fig. 4, the angular difference θ is determined by the starting values v_{xL} and v_{yL}, and the ending values v_{xR} and v_{yR} of linear trends of the compares segments, as well as the segment duration t_i ($t_i = t_{iR} - t_{iL}$). The maximum angular difference $\theta_{\max}$ is determined by the maximum and minimum values of the approximated PLRs of the compared sequences x_L and y'_L, the starting values of the compared linear trends v_{xL} and v_{yL} as well as the segment duration t_i. It can be inferred that the segment trend similarity $TS(x_i, y_i) \in [-1, 1]$ ($TS(x_i, y_i) \in [0, 1]$ for segment trends in a same direction, $TS(x_i, y_i) \in [-1, 0)$ for segment trends in different directions). The segment trends are more similar when $TS(x_i, y_i)$ is close to 1 and they are more reverse when $TS(x_i, y_i)$ is close to -1. $TS(x_i, y_i) > 0$ indicates the compared segment trends are in the same direction while $TS(x_i, y_i) < 0$ indicates the segment trends are in different directions.

Segment composite similarity

The segment composite similarity $CS(x_i, y_i)$ combines both segment distance and trend similarities using the weighted sum method which assigns different weights according to the importance of each metric, as shown in Eq. (2).

$$CS(x_i, y_i) = w_1 \cdot DS(x_i, y_i) + w_2 \cdot TS(x_i, y_i) \tag{2}$$

The existing evaluation standards and studies primarily evaluate distance metrics (e.g., MAD, RMSD) between the estimations and references. In this study, we emphasized the importance of trend tracking in cuffless BP monitoring. Therefore, we assign the weight for distance similarity $w_1 = 0.6$ and the weight for trend similarity $w_2 = 0.4$. And these weights can be adjusted according to different evaluation priorities.

Combining the segment similarity measures (global evaluation)

Inspired by the method used in[51], a temporal-normalization method which combines the segment similarity measures based on the proportion of the segment duration was applied for this study. The temporally normalized distance, trend and composite similarity of the compared reference and estimation of BPs are therefore defined and calculated as shown in Table 5.

As shown in Table 5, $DS(x_i, y_i)$, $TS(x_i, y_i)$ and $CS(x_i, y_i)$ are the segment distance, trend and composite similarity respectively. t_i is the duration of the *i*-th segment, t_D is the duration of the compared sequences, *K* is the number of segments. Therefore, $\sum_{i=1}^{K} \frac{t_i}{t_D} = 1$ which means the temporal normalization method is similar to the weighted sum method and the weight of each segment result contributes to the final result is determined by the segment duration, indicating the longer the segment is, the more ratios it contributes to the final result. Also, it can be inferred that $\text{TN\_DS}(x, y) \in [-1, 1]$, $\text{TN\_TS}(x, y) \in [-1, 1]$ and $\text{TN\_CS}(x, y) \in [-1, 1]$. The segment distance, trend and composite similarity metrics, as well as their temporally normalized results for whole sequence comparison of the example shown in Fig. 5 are illustrated in Table 6.

The results illustrated in Table 6 indicate PCC and CCC can not properly evaluate trend similarity between the reference and estimated BPs. For example, the trends of the 5th segment are almost parallel and the MAD between the estimations and references is only 0.44 mmHg which are accurately evaluated by new proposed distance and trend similarity metrics, whereas the PCC and CCC are 0.601 and 0.096. It can also be observed in Fig. 5 that the trends of the 2nd segment are more distinct than the 1st segment with the proposed trend similarity of -0.68 and -0.14 respectively. However, the PCC values for these two segments are identical.

Interpretation of the proposed metrics

The proposed segment distance, trend similarity metrics and their composite metric, as well as their temporally-normalized versions for whole sequence evaluation, are normalized metrics (defined in[53]) and they have a same range of $[-1, 1]$.

The estimation is more reliable when the distance similarity $DS(x, y) > 0$ (also applicable to segment distance similarity $DS(x_i, y_i)$) and it is more accurate when *DS*(*x*, *y*) is close to 1. The estimation is more unreliable when $DS(x, y) < 0$ and it is more inaccurate when *DS*(*x*, *y*) is close to -1. The estimated and the reference segment BP trends are in the same direction when the segment trend similarity $TS(x_i, y_i) > 0$ and they are more similar when $TS(x_i, y_i)$ is close to 1. The estimated and the reference segment BP trends are in opposite directions when the segment trend similarity $TS(x_i, y_i) < 0$ and they are more reverse when $TS(x_i, y_i)$ is close to -1. The estimated BP pattern are more similar to the reference BP pattern when *TS*(*x*, *y*) is close to 1 and the majority of the estimated and the reference BP trends are in the same direction when $TS(x, y) > 0$ and vice versa.

Since the composite similarity metric *CS*(*x*, *y*) combines distance *DS*(*x*, *y*) and trend *TS*(*x*, *y*) similarity metrics with weighted sum, the estimation is more accurate in terms of both distance and trend when *CS*(*x*, *y*) is close to 1, the estimation is more different to the reference in terms of distance and/or trend when *CS*(*x*, *y*) is close to -1. Therefore, a BP estimation model is expected to have the composite similarity metric *CS*(*x*, *y*) that is as close to 1 as possible and we suggest that a BP estimation method/device can be considered capable for tracking BP changes if the proposed metrics TN_DS≥0, TN_TS≥0 and TN_CS≥0. Detailed interpretation of the proposed metrics with a set of simulations can be found in Section S8.

	x_{L_i} and y'_{L_i} in the same direction slope (x_{L_i})× slope (y'_{L_i})≥ 0		x_{L_i} and y'_{L_i} in different directions slope (x_{L_i})× slope (y'_{L_i})< 0	
	Slope(x_{L_i})≥ 0, slope(y'_{L_i})≥ 0 (Fig. 4a)	Slope(x_{L_i})≤ 0, slope(y'_{L_i})≤ 0 (Fig. 4b)	Slope(x_{L_i})> 0, slope(y'_{L_i})< 0 (Fig. 4c)	Slope(x_{L_i})< 0, slope(y'_{L_i})> 0 (Fig. 4d)
$TS(x_i, y_i)$	$1 - \frac{\theta}{\theta_{\max}}$		$-\frac{\theta}{\theta_{\max}}$	
Where				
θ	$\lvert \arctan(\frac{v_{xR}-v_{xL}}{t_{iR}-t_{iL}}) - \arctan(\frac{v_{yR}-v_{yL}}{t_{iR}-t_{iL}}) \rvert$		$\arctan(\frac{\lvert v_{xR}-v_{xL} \rvert}{t_{iR}-t_{iL}}) + \arctan(\frac{\lvert v_{yR}-v_{yL} \rvert}{t_{iR}-t_{iL}})$	
$\theta_{\max}$	$\arctan(\frac{\max(x_L,y_L)-\min(v_{xL},v_{yL})}{t_{iR}-t_{iL}})$	$\arctan(\frac{\max(v_{xL},v_{yL})-\min(x_L,y_L)}{t_{iR}-t_{iL}})$	$\arctan(\frac{\max(x_L,y_L)-v_{xL}}{t_{iR}-t_{iL}}) + \arctan(\frac{v_{yL}-\min(x_L,y_L)}{t_{iR}-t_{iL}})$	$\arctan(\frac{\max(x_L,y_L)-v_{yL}}{t_{iR}-t_{iL}}) + \arctan(\frac{v_{xL}-\min(x_L,y_L)}{t_{iR}-t_{iL}})$

Table 4. Segment trend similarity for different scenarios.

Temporally-normalized similarity of distance (TN_DS)	$\mathrm{TN\_DS}(x,y) = \frac{1}{t_D}\sum_{i=1}^{K} DS(x_i,y_i)\cdot t_i$
Temporally-normalized similarity of trend (TN_TS)	$\mathrm{TN\_TS}(x,y) = \frac{1}{t_D}\sum_{i=1}^{K} TS(x_i,y_i)\cdot t_i$
Temporally-normalized composite similarity of distance and trend (TN_CS)	$\mathrm{TN\_CS}(x_i,y_i) = \frac{1}{t_D}\sum_{i=1}^{K} CS(x_i,y_i)\cdot t_i$

Table 5. Equations of calculating temporally-normalized similarity metrics.

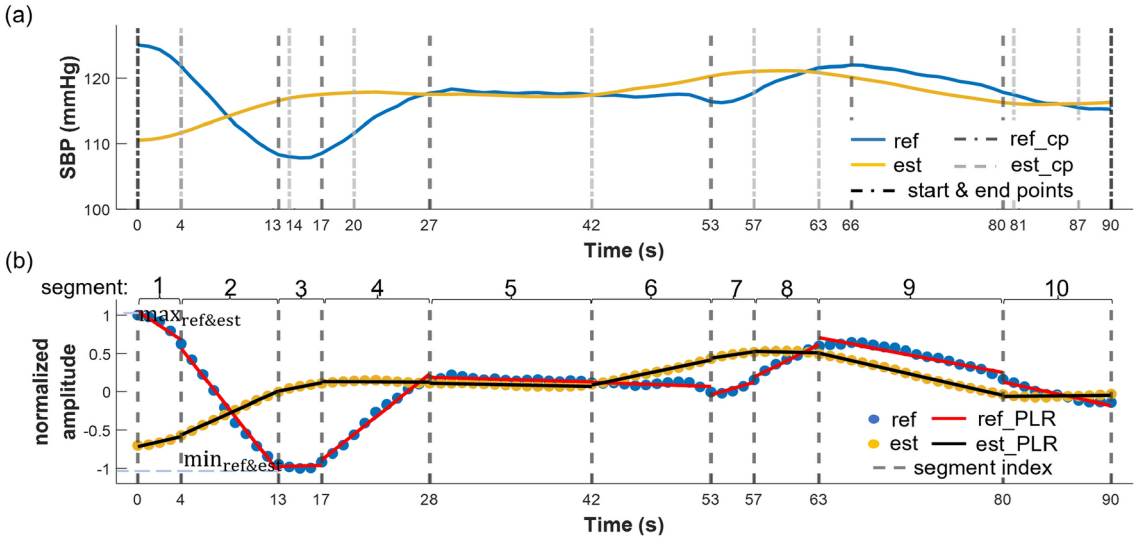


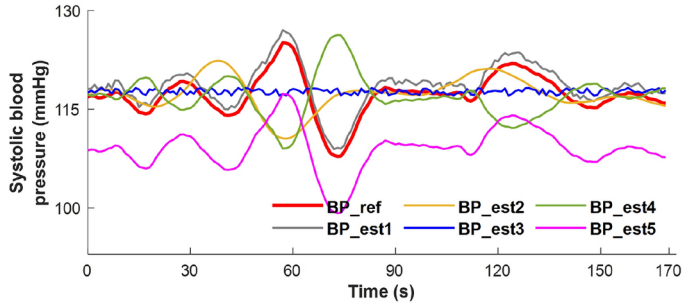
Fig. 5. A demonstration of (**a**) segmenting the compared reference and estimated BPs using BEAST method, the reference BP is divided into 8 segments while the estimated BP is divided into 9 segments ($t = 4$ is a change point for both reference and estimation); (**b**) the equal segment number (ESN) will re-process the change points and re-divide the compared sequences such that the compared sequences have the same number of segments (e.g., the reference and estimated BPs are re-processed to 10 segments), and the least squares are used to approximate the linear trend for each segment.

	Segment (local)										Whole sequence (global)
	1	2	3	4	5	6	7	8	9	10	
Duration [Start,end]	4s [0, 4)	9s [4, 13)	4s [13, 17)	10s [17, 27)	15s [27, 42)	11s [42, 53)	4s [53, 57)	6s [57, 63)	17s [63, 80)	10s [80, 90]	77s [0, 77]
MAD (mmHg)	12.87	5.28	9.00	4.13	0.44	1.51	3.96	1.39	1.99	0.75	2.96
$\mathrm{PCC}(x_i,y_i)$	−0.998	−0.998	−0.012	−0.082	0.601	−0.532	0.788	−0.465	0.949	−0.126	−0.016
$\mathrm{CCC}(x_i,y_i)$	−0.007	−0.625	0	−0.003	0.096	−0.113	0.019	−0.049	0.432	−0.030	−0.015
$DS(x_i,y_i)$	−1.000	0.240	−0.999	0.907	1.000	0.999	0.935	0.999	0.999	1	0.73
$TS(x_i,y_i)$	−0.14	−0.68	0.95	−0.36	0.98	−0.18	0.93	−0.19	0.94	−0.14	0.26
$CS(x_i,y_i)$	−0.66	−0.13	−0.22	0.40	0.99	0.53	0.93	0.53	0.98	0.54	0.54

Table 6. Segment similarity measures of the example shown in Fig. 5 proposed by existing works and this study and their temporally-normalized results for whole sequences comparison.

Experiments and results

To demonstrate the performance of the proposed distance, trend and composite similarity metrics, these metrics are firstly applied to compare a set of different simulated BP estimations to the reference BP for feasibility validation. Then, these new metrics are applied to compare the estimated BPs using two SOTA models to the reference BPs of a real world dataset.



BP_{est1}	$BP_{est1} = 1.05 \cdot BP_{ref} - 5 +$ Gaussian noise
BP_{est2}	*1)* $BP_{est21} = MA(BP_{ref}, 15)$, *2)* $BP_{est2} = resample(BP_{est21}(25{:}end), length(BP_{ref}))$
BP_{est3}	$BP_{est3} = mean(BP_{ref}) +$ Gaussian noise
BP_{est4}	$BP_{est4} = -BP_{ref} + 2 \cdot mean(BP_{ref})$
BP_{est5}	$BP_{est5} = 1.05 \cdot BP_{ref} - 14$

MA: moving average, MA(x,k) where x is the sequence and k is the size of sliding window

Fig. 6. Waveforms of the reference SBP measurement (BP_{ref}) and five simulated SBP estimations (BP_{est1} to BP_{est5}), and relations between the simulated SBPs and reference SBP.

	Conventional metrics					Composite measure	Temporally normalized metrics (ours)		
	$MD \pm SDD^*$	$MAD \pm SDAD^*$	$RMSD^*$	PCC	CCC	GCSM	TN_DS	TN_TS	TN_CS
BP_{est1}&BP_{ref}	1.36 ± 0.34	1.36 ± 0.34	1.40	0.996	0.914	0.91	0.999	0.98	0.99
BP_{est2}&BP_{ref}	0.08 ± 4.45	2.85 ± 3.41	4.43	-0.23	-0.22	-0.19	0.56	0.17	0.40
BP_{est3}&BP_{ref}	0.51 ± 3.16	2.27 ± 2.26	3.20	-0.02	-0.004	-0.012	0.81	0.34	0.62
BP_{est4}&BP_{ref}	-0.28 ± 6.28	4.30 ± 4.57	6.27	-1	-0.996	-1	0.34	-0.46	0.02
BP_{est5}&BP_{ref}	-8.14 ± 0.16	8.14 ± 0.16	8.14	1	0.24	0.68	-0.994	0.98	-0.20

Table 7. Conventional distance metrics, correlation coefficients, composite similarity measure GCSM and the proposed similarity metrics calculated between simulated BP estimations and reference BP. *unit: mmHg.

Applying the proposed new metrics to simulated data

In this section, a synthetic dataset spanning a 3-minute interval was generated, simulating BP estimations using reference experimental data characterized by substantial variations within that time frame. Consequently, the dataset includes waveforms with varying degrees of fidelity such that various conventional metrics and the proposed metrics can be assessed in tracking BP changes. The waveforms of the reference BP and simulated BP estimations and their relations are illustrated in Fig. 6. BP_{est1} and BP_{est5} have almost identical pattern to BP_{ref} but with different distances. BP_{est2} simulates a random estimation of BP_{ref}. BP_{est3} simulates an estimation model that is unable to track any BP change and just estimates the average BP level, and BP_{est4} simulates a reverse estimation of BP_{ref}.

The conventional distance metrics (MD, MAD, RMSD), correlation coefficients (PCC and CCC), a composite similarity measure GCSM[44], and the proposed similarity metrics between different simulated BP estimations and the reference BP are shown in Table 7.

It can be observed from Table 7 that conventional distance metrics can only measure the dissimilarity of pairwise values for the compared sequences. Though PCC can indicates the pattern similarities between the simulated BP estimations and the reference. For example, BP_{est5} and BP_{est1} have almost identical morphology to BP_{ref} and therefore PCC=1, PCC=-1 for BP_{est4} indicates the estimated BP has a reverse morphology to BP_{ref} and $PCC \approx 0$ for BP_{est3} indicates the estimation model does not track any BP changes. The GCSM, which composes similarities of mean, standard deviation and PCC, demonstrates its potential in assessing both distance and trend similarities for most of the scenarios, except for BP_{est5}. As previously stated, the sensitivity of $s_1(x, y)$ and $s_2(x, y)$ of GCSM (see "A composite score to represent similarity in mean, variance and correlation coefficient") is significantly affected by outliers or instances where the disparity between the global maximum and global minimum values ($\max_{x,y} - \min_{x,y}$ in Table 3) is substantial. Consequently, $s_1(x, y)$ and $s_2(x, y)$ when comparing BP_{est5} and BP_{ref} will still approximate to 1 even though the MAD is 8.14 mmHg, which is a considerable large error for cuffless BP estimation. CCC and GCSM will exhibit more limitations in cases when the compared time series are long and noisy which can be observed from results obtained from DailA BP in "Applying the proposed new metrics to results of two SOTA models". The new similarity metrics are proposed to address the drawbacks of the aforementioned metrics, and the results from this simulated dataset validated the proposed methods can accurately quantify both distance and trend similarity for cuffless BP studies.

Applying the proposed new metrics to results of two SOTA models

The wearable physiological and BP measurements during activities of daily living dataset[54] (from now on referred as DailA BP) will be used to train and evaluate two SOTA cuffless BP estimation models with proposed metrics. DailA BP includes approximately 6.5 hours' physiological waveforms (ECG, PPG and BP) from 5 healthy subjects (4 males, age (28±6.6) yrs) performing daily activities and several activities such as Valsalva Maneuver, walking and hand gripping were performed to induce BP changes (see Section S2 and Fig. S1).

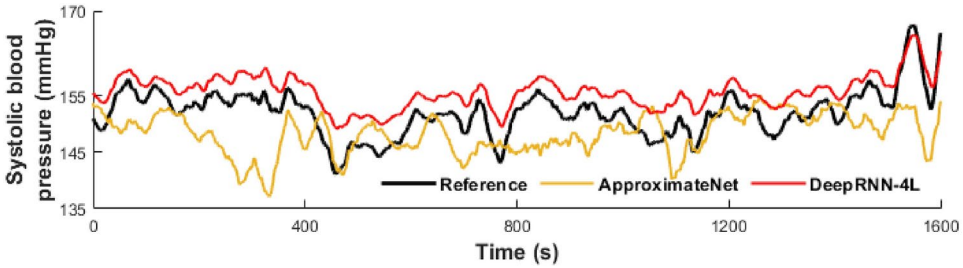


Fig. 7. An approximate 30 minutes' example of reference SBPs and estimated SBPs using SOTA models (BP variations might be caused by daily activities, such us standing up and walking), the distance, trend and composite similarities between estimated SBPs using ApproximateNet and reference SBPs are 0.37, 0.23 and 0.32 respectively, and 0.81, 0.81 and 0.81 between estimated BPs using DeepRNN-4L and reference BPs respectively (the composite similarities with different combinations of w_1 and w_2 are demonstrated in Sect. S9).

		Conventional metrics					Composite measure	Temporally normalized metrics (ours)		
		MD ± SDD*	MAD ± SDAD*	RMSD*	PCC	CCC	GCSM	TN_DS	TN_TS	TN_CS
ApproximateNet	SBP	-5.82 ± 11.77	9.88 ± 8.65	13.13	0.68	0.62	0.64	-0.22	0.36	0.015
	DBP	-1.70 ± 8.04	6.50 ± 5.02	8.22	0.65	0.62	0.61	-0.001	0.34	0.13
DeepRNN-4L	SBP	0.21 ± 5.28	4.13 ± 3.29	5.29	0.94	0.93	0.92	0.87	0.70	0.80
	DBP	0.77 ± 5.25	4.15 ± 3.30	5.30	0.75	0.73	0.72	0.74	0.55	0.66

Table 8. Conventional distance metrics, correlation coefficients, composite similarity measure GCSM and the proposed similarity metrics calculated between the estimated BPs and reference BPs using SOTA models from DailA BP dataset. *Unit: mmHg.

Two SOTA models , which differ in BP estimation errors and trend tracking, were selected to assess the feasibility of the proposed metrics. ApproximateNet[12] which is a deeply supervised one-dimensional U-Net model, was proposed to approximate BP waveform from a simultaneous PPG waveform. It was originally trained and evaluated with 127,260 subsets of simultaneous PPG and BP waveforms from UCI BP[55]. The reported MAD±SDAD for SBP and DBP estimation is (6.42±9.52) mmHg and (3.58±5.80) mmHg respectively[12]. DeepRNN-4L[56], which consists of a bidirectional LSTM layer and four forward LSTM layers with residual connections, was proposed to estimate SBP and DBP using 7 features extracted from ECG and PPG in a time series analysis fashion. DeepRNN-4L was originally trained and evaluated on the static continuous BP dataset which includes simultaneous ECG, PPG and BP data from 84 healthy subjects (51 males and 33 females) while stationary, and the reported RMSD of SBP and DBP is 3.75 and 2.43 mmHg respectively[56]. Different training strategies were applied for ApproximateNet and DeepRNN-4L models (details of implementation, model parameters and training processes are described in Section S5), also due to the fact that DeepRNN-4L model considers temporal dependency. Therefore, these two models generate different estimations on the same data in terms of both distance and pattern tracking accuracy. A proportion (approximate 30 minutes) of reference SBPs and estimated SBPs using two SOTA models are shown in Fig. 7.

The conventional distance metrics (MD, MAD, RMSD), correlation coefficients (PCC and CCC), a composite similarity measure GCSM, and the proposed similarity metrics between BPs estimated by two SOTA models and reference BPs from DailA BP dataset are shown in Table 8.

As shown in Table 8 that DeepRNN-4L outperforms ApproximateNet in estimating BP using DailA BP dataset by comparing distance metrics and correlation coefficients. But these conventional metrics can not indicate the performance of these estimation models in BP change tracking. The GCSM results almost align with PCC, this is because the denominators of the $s_1(x, y)$ and $s_2(x, y)$ can be much larger than their numerators ($\max_{x,y} - \min_{x,y} \gg |\mu_x - \mu_y|$ or $(\max_{x,y} - \min_{x,y})/2 \gg |\sigma_x - \sigma_y|$) and therefore $s_1(x, y)$ or $s_2(x, y) \approx 1$ for long and noisy sequences. Eventually, GCSM is mostly determined by $s_3(x, y)$ which is PCC of the compared sequences. The aforementioned issues are tackled by the proposed methods by the utilization of PLR which simplify noisy and high dimensional time series with a number of line segments. This approach leads to improved accuracy in quantifying trend similarity. The utilization of temporal normalization also effectively tackles the problem associated with evaluation strategy by providing varying weights to the final outcome based on the duration of each segment. The distance similarity TN_DS of ApproximateNet is -0.22 and -0.001 for SBP and DBP respectively, indicating the estimations of this method more likely dissatisfy IEEE Standard. DeepRNN-4L outperforms ApproximateNet in terms of both distance and trend tracking, which is also indicated by the composite similarity TN_CS.

Conclusion

Continuous cuffless BP tracking has become a popular research problem recently and has shown promising results. However, current evaluation standards which primarily use pairwise distance metrics are not robust and comprehensive. In this paper, we firstly reviewed several commonly used BP evaluation standards and then introduced BP patterns and the importance of accurate tracking of BP changes. The conventional distance metrics, correlation coefficients which are commonly used by existing studies, a composite similarity metric GCSM were reviewed and their drawbacks in terms of BP pattern tracking were described and demonstrated with examples. Finally, we proposed a comprehensive framework with new distance, trend and composite similarity metrics with temporal normalization, for continuous cuffless BP evaluation.

The results from the simulated dataset and the re-implementation of two SOTA models using a realistic dataset DailA BP verified the feasibility of the proposed similarity metrics for cuffless BP evaluation. Our results demonstrated that the composite similarity metric has advantages in evaluating both distance and trend similarities for long and noisy time series, comparing to conventional metrics and GCSM. Additionally, the proposed metrics are normalized metrics and range from -1 to 1, making them intuitively interpretable, similar to well-known correlation coefficients. The proposed methods and metrics can also be used to benchmark different algorithms for cuffless BP estimation, diagnose and improve the estimation accuracy by analyzing proposed metrics (see Section S7), facilitating the adoption of this new technique in real-world applications. The future work will adapt the proposed similarity metrics as customizable loss function to improve the performance of estimation models in tracking BP pattern.

Data availibility

The datasets generated and analysed during the current study, as well as the MATLAB implementation of the proposed methods with examples are available in the GitHub repository, https://github.com/shanhe0426/metrics-for-cuffless-blood-pressure.git.

Received: 30 May 2024; Accepted: 21 October 2024
Published online: 11 November 2024

References

1. Blood pressure. https://en.wikipedia.org/wiki/Blood_pressure.
2. Frieden, T. Taming the world's leading killer: High blood pressure. https://www.cnn.com/2021/05/30/health/high-blood-pressure-frieden/index.html.
3. Perloff, D. et al. Human blood pressure determination by sphygmomanometry. *Circulation* **88**, 2460–2470 (1993).
4. Alpert, B. S., Quinn, D. & Gallick, D. Oscillometric blood pressure: A review for clinicians. *J. Am. Soc. Hypertens.* **8**, 930–938 (2014).
5. Mukkamala, R., Stergiou, G. S. & Avolio, A. P. Cuffless blood pressure measurement. *Annu. Rev. Biomed. Eng.* **24**, 203–230 (2022).
6. Mukkamala, R. Blood pressure with a click of a camera? *Circ. Cardiovasc. Imaging* **12**, e009531 (2019).
7. Seo, J., Pietrangelo, S. J., Lee, H.-S. & Sodini, C. G. Noninvasive arterial blood pressure waveform monitoring using two-element ultrasound system. *IEEE Trans. Ultrason. Ferroelectr. Freq. Control* **62**, 776–784 (2015).
8. Beulen, B. W. et al. Toward noninvasive blood pressure assessment in arteries by using ultrasound. *Ultrasound Med. Biol.* **37**, 788–797 (2011).
9. Zheng, Y., Poon, C. C., Yan, B. P. & Lau, J. Y. Pulse arrival time based cuff-less and 24-h wearable blood pressure monitoring and its diagnostic value in hypertension. *J. Med. Syst.* **40**, 1–11 (2016).
10. Hsu, Y.-C., Li, Y.-H., Chang, C.-C. & Harfiya, L. N. Generalized deep neural network model for cuffless blood pressure estimation with photoplethysmogram signal only. *Sensors* **20**, 5668 (2020).
11. Martínez, G. et al. Can photoplethysmography replace arterial blood pressure in the assessment of blood pressure?. *J. Clin. Med.* **7**, 316 (2018).
12. Ibtehaz, N. et al. Ppg2abp: Translating photoplethysmogram (ppg) signals to arterial blood pressure (abp) waveforms. *Bioengineering* **9**, 692 (2022).
13. O'Brien, E. et al. The British hypertension society protocol for the evaluation of blood pressure measuring devices. *J. Hypertens.* **11**, S43–S62 (1993).
14. Association for the Advancement of Medical Instrumentation and others. American National Standard. Manual, electronic or automated sphygmomanometers. *ANSI/AAMI SP10-2002/A1* (2003).
15. O'Brien, E. et al. European Society of Hypertension International Protocol Revision 2010 for the validation of blood pressure measuring devices in adults. *Blood Press. Monit.* **15**, 23–38 (2010).
16. Stergiou, G. S. et al. A universal standard for the validation of blood pressure measuring devices: Association for the advancement of medical instrumentation/european society of hypertension/international organization for standardization (aami/esh/iso) collaboration statement. *Hypertension* **71**, 368–374 (2018).
17. El-Hajj, C. & Kyriacou, P. A. Deep learning models for cuffless blood pressure monitoring from ppg signals using attention mechanism. *Biomed. Signal Process. Control* **65**, 102301 (2021).
18. Stergiou, G. S. et al. Cuffless blood pressure measuring devices: review and statement by the European Society of Hypertension Working Group on blood pressure monitoring and cardiovascular variability. *J. Hypertens.* **40**, 1449–1460 (2022).
19. IEEE standard for wearable cuffless blood pressure measuring devices. In *IEEE Std 1708-2014*. 1–38. https://doi.org/10.1109/IEEESTD.2014.6882122 (2014).
20. IEEE standard for wearable, cuffless blood pressure measuring devices—amendment 1. In *IEEE Std 1708a-2019 (Amendment to IEEE Std 1708-2014)*. 1–35. https://doi.org/10.1109/IEEESTD.2019.8859685 (2019).
21. Non-invasive sphygmomanometers—part 3: Clinical investigation of continuous automated measurement type. *ISO 81060-3:2022*. 1–36 (2022).
22. Stergiou, G. S. et al. European society of hypertension recommendations for the validation of cuffless blood pressure measuring devices: European Society of Hypertension Working Group on blood pressure monitoring and cardiovascular variability. *J. Hypertens.* 10–1097 (2023).
23. Weber, T. et al. Twenty-four-hour central (aortic) systolic blood pressure: reference values and dipping patterns in untreated individuals. *Hypertension* **79**, 251–260 (2022).
24. Schutte, A. E., Kollias, A. & Stergiou, G. S. Blood pressure and its variability: Classic and novel measurement techniques. *Nat. Rev. Cardiol.* **19**, 643–654 (2022).

25. Di Raimondo, D., Musiari, G. & Pinto, A. Nocturnal blood pressure patterns and cardiac damage: There is still much to learn. *Hypertens. Res.* **43**, 246–248 (2020).
26. Parati, G., Ochoa, J. E., Lombardi, C. & Bilo, G. Assessment and management of blood-pressure variability. *Nat. Rev. Cardiol.* **10**, 143–155 (2013).
27. Parati, G., Torlasco, C., Pengo, M., Bilo, G. & Ochoa, J. E. Blood pressure variability: Its relevance for cardiovascular homeostasis and cardiovascular diseases. *Hypertens. Res.* **43**, 609–620 (2020).
28. O'Brien, E., Kario, K., Staessen, J. A., de la Sierra, A. & Ohkubo, T. Patterns of ambulatory blood pressure: Clinical relevance and application. *J. Clin. Hypertens.* **20**, 1112–1115 (2018).
29. Pal, A. et al. Beat-to-beat blood pressure variability in patients with obstructive sleep apnea. *J. Clin. Sleep Med.* **17**, 381–392 (2021).
30. Phillips, C. L. & O'Driscoll, D. M. Hypertension and obstructive sleep apnea. *Nat. Sci. Sleep*. 43–52 (2013).
31. Dani, M. et al. Orthostatic hypotension in older people: Considerations, diagnosis and management. *Clin. Med.* **21**, e275 (2021).
32. Freeman, R. et al. Consensus statement on the definition of orthostatic hypotension, neurally mediated syncope and the postural tachycardia syndrome. *Auton. Neurosci.* **161**, 46–48 (2011).
33. Ohkubo, T. et al. Prediction of stroke by ambulatory blood pressure monitoring versus screening blood pressure measurements in a general population: The Ohasama study. *J. Hypertens.* **18**, 847–854 (2000).
34. Dolan, E. et al. Superiority of ambulatory over clinic blood pressure measurement in predicting mortality: The Dublin outcome study. *Hypertension* **46**, 156–161 (2005).
35. Trinder, J. et al. Autonomic activity during human sleep as a function of time and sleep stage. *J. Sleep Res.* **10**, 253–264 (2001).
36. Sakoe, H. & Chiba, S. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Trans. Acoust. Speech Signal Process.* **26**, 43–49 (1978).
37. Alt, H. & Godau, M. Computing the Fréchet distance between two polygonal curves. *Int. J. Comput. Geom. Appl.* **5**, 75–91 (1995).
38. Huttenlocher, D. P., Klanderman, G. A. & Rucklidge, W. J. Comparing images using the Hausdorff distance. *IEEE Trans. Pattern Anal. Mach. Intell.* **15**, 850–863 (1993).
39. Chowdhury, M. H. et al. Estimating blood pressure from the photoplethysmogram signal and demographic features using machine learning techniques. *Sensors* **20**, 3127 (2020).
40. Lawrence, I. & Lin, K. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 255–268 (1989).
41. Mukaka, M. Statistics corner: A guide to appropriate use of correlation in medical research. *Malawi Med. J.* **24**, 69–71 (2012).
42. Overholser, B. R. & Sowinski, K. M. Biostatistics primer: Part 2. *Nutr. Clin. Pract.* **23**, 76–84 (2008).
43. McBride, G. et al. A proposal for strength-of-agreement criteria for Lin's concordance correlation coefficient. In *NIWA Client Report: HAM2005-062*. Vol. 45. 307–310 (2005).
44. Liu, Y., Kim, K. S., Beresford, R. M. & Fleisher, D. H. A generic composite measure of similarity between geospatial variables. *Ecol. Inform.* **60**, 101169 (2020).
45. Cassisi, C. et al. Similarity measures and dimensionality reduction techniques for time series data mining. In *Advances in Data Mining Knowledge Discovery and Applications*. 71–96 (2012).
46. Pratt, K. B. & Fink, E. Search for patterns in compressed time series. *Int. J. Image Graph.* **2**, 89–106 (2002).
47. Keogh, E., Chu, S., Hart, D. & Pazzani, M. Segmenting time series: A survey and novel approach. data mining in time series databases. *World Sci.* **57**, 1–22 (2004).
48. Ramer, U. An iterative procedure for the polygonal approximation of plane curves. *Comput. Graph. Image Process.* **1**, 244–256 (1972).
49. Keogh, E. Fast similarity search in the presence of longitudinal scaling in time series databases. In *Proceedings Ninth IEEE International Conference on Tools with Artificial Intelligence*. 578–584 (IEEE, 1997).
50. Khatami, M. & Akbarzadeh, F. Algorithms for segmenting time series. *Global Anal. Discrete Math.* **3**, 65–73 (2018).
51. WangDa, D. et al. Pattern distance of time series. In *WIT Transactions on Information and Communication Technologies*. Vol. 29 (2003).
52. Zhao, K. et al. Detecting change-point, trend, and seasonality in satellite time series data to track abrupt changes and nonlinear dynamics: A bayesian ensemble algorithm. *Remote Sens. Environ.* **232**, 111181 (2019).
53. Chen, S., Ma, B. & Zhang, K. On the similarity metric and the distance metric. *Theor. Comput. Sci.* **410**, 2365–2376 (2009).
54. Landry, C., Hedge, E. T., Hughson, R. L., Peterson, S. D. & Arami, A. Wearable physiological and blood pressure measurements during activities of daily living. https://doi.org/10.21227/wysp-gt69 (2021).
55. Kachuee, M., Kiani, M., Mohammadzade, H. & Shabany, M. Cuff-less blood pressure estimation. In *UCI Machine Learning Repository*. https://doi.org/10.24432/C5B602 (2015).
56. Su, P. et al. Long-term blood pressure prediction with deep recurrent neural networks. In *2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*. 323–328 (IEEE, 2018).

Author contributions

Shan He: conceptualization, methodology, programming, and writing - original draft preparation. Miodrag Bolić: supervision, project administration and writing - review and edit. All authors have read and agreed to the published version of manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-77171-6.

Correspondence and requests for materials should be addressed to S.H.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.