scientific reports

OPEN

Examining the challenges of blood pressure estimation via photoplethysmogram

Suril Mehta[1]✉, Nipun Kwatra[1], Mohit Jain[1] & Daniel McDuff[2]

The use of observed wearable sensor data (e.g., photoplethysmograms [PPG]) to infer health measures (e.g., glucose level or blood pressure) is a very active area of research. Such technology can have a significant impact on health screening, chronic disease management and remote monitoring. A common approach is to collect sensor data and corresponding labels from a clinical grade device (e.g., blood pressure cuff) and train deep learning models to map one to the other. Although well intentioned, this approach often ignores a principled analysis of whether the input sensor data have enough information to predict the desired metric. We analyze the task of predicting blood pressure from PPG pulse wave analysis. Our review of the prior work reveals that many papers fall prey to data leakage and unrealistic constraints on the task and preprocessing steps. We propose a set of tools to help determine if the input signal in question (e.g., PPG) is indeed a good predictor of the desired label (e.g., blood pressure). Using our proposed tools, we found that blood pressure prediction using PPG has a high multi-valued mapping factor of 33.2% and low mutual information of 9.8%. In comparison, heart rate prediction using PPG, a well-established task, has a very low multi-valued mapping factor of 0.75% and high mutual information of 87.7%. We argue that these results provide a more realistic representation of the current progress toward the goal of wearable blood pressure measurement via PPG pulse wave analysis. For code, see our project page: https://github.com/lirus7/PPG-BP-Analysis

The COVID-19 pandemic has highlighted the acute need for technology to support remote health care[1,2]. Consultancy McKinsey[3] reported a 40-fold increase in the use of telehealth services and a 40% increase in consumer interest in virtual health solutions when compared to pre-COVID-19 statistics. To provide an example, the ability to estimate vital signs from sensors available in smartphones and wearable devices could have a significant impact on the effective management of diseases (e.g., COVID-19, hypertension, diabetes). Frequent measurement of physiological parameters can help in managing medication dosages and understanding the effects of lifestyle changes on health.

The estimation of vital signs traditionally relies on customized sensors that measure physical or chemical properties of the body. For example, digital sphygmomanometers use sensors to measure the oscillations in the arteries to quantify blood pressure. Although accurate, such medical devices are far from ubiquitous, often are not easy to access and are uncomfortable to use for extended periods of time. An alternative approach, promoted by the field of ubiquitous computing is to leverage sensors already present in every day devices for estimating health parameters. For example, heart rate can be measured using a smartphone camera by analyzing subtle changes in skin color as the heart pumps blood around the body[4,5]. This technology is now available on billions of devices, Google Fit (https://www.google.com/fit/). Recent work has presented proof-of-concept measurement of oxygen saturation[6], blood pressure[7], and hemoglobin levels[8] via smartphones.

Existing research work can be broadly divided into two categories: (1) approaches that are developed from first principles to imitate an established medical method for measurement or diagnosis[9,10], and (2) approaches where input (sensor) data and corresponding gold-standard data are collected using a medical grade device and machine learning models are trained to discover a relationship between the input and output[11,12]. In this paper, we focus on the latter category. Although well-intentioned, such data-driven approaches ignore a principled analysis of whether the input data have the necessary information to predict the desired health measure. As a result, numerous human and compute hours are wasted in developing and training deep learning models for prediction tasks that may be ill-posed or not feasible.

We consider the task of predicting blood pressure (BP) non-invasively. Blood Pressure is the pressure applied on arterial walls as the blood circulates through the body. It depends on multiple factors, including blood volume,

[1]Microsoft Research, Bengaluru, India. [2]Microsoft Research, Redmond, USA. ✉email: suril15104@iiitd.ac.in

blood viscosity, and stiffness of blood vessels. Abnormally high or low blood pressure can result in heart attack, stroke, and diabetes[13,14] thus it is recommended to measure BP frequently.

The methods to measure blood pressure non-invasively can be broadly categorized into two approaches: (i) The pulse transit time (PTT) method[15–17] is a popular, non-invasive technique for measuring blood pressure based on the time delay for a pressure wave to travel between proximal and distal arterial sites. The PTT approach has strong theoretical underpinnings based on the Bramwell-Hill equation[18], which relates PTT to pulse wave velocity and arterial compliance. The Wesseling model captures the relationship between arterial compliance and blood pressure[19]. However, it is important to note that, PTT can change independently of BP due to factors such as aging-induced arteriosclerosis, and smooth muscle contraction. Hence, it needs to be calibrated from time to time. (ii) Pulse Wave Analysis (PWA) is a method used to estimate blood pressure (BP) by extracting features from an arterial waveform. This is typically performed using a photoplethysmography (PPG) waveform. PPG is an optical signal obtained by illuminating the skin (common sites are the finger, earlobe, or toe[20]) with an LED and measuring the amount of transmitted, or reflected, light using a photodiode. PPG detects blood volume changes in the microvascular bed of tissue, as the blood volume directly impacts the amount of light transmitted/reflected. Unlike PTT, PWA has weaker theoretical underpinnings as the small arteries interrogated by PPG are viscoelastic[15]. Calibration is invariably necessary for PWA analysis methods to obtain reasonable results.

In this study, we concentrate on PWA measurement of BP. This method is beneficial because it only requires the use of a single sensor making it a more accessible solution. Predicting BP by analyzing PPG waveforms is an active area of research[7,21–25] and is already used in consumer products (https://www.samsung.com/global/galaxy/what-is/blood-pressure/). However, we should note that "*while these methods (PTT and PWA) have been extensively studied and cuff-calibrated devices are now on the market, there is no compelling proof in the public domain indicating that they can accurately track intra-individual BP changes*"[20,26]. Therefore, although the features extracted from the PPG signal correlate with blood pressure, the signal's adequacy for accurately predicting blood pressure remains unclear.

The discrepancy between recent research[27–29] claiming promising results on evaluation benchmarks for blood pressure, and other observational studies[20,26] which indicate a lack of a concrete theory to measure blood pressure using PPG signals via PWA, raises important questions. To help resolve this apparent contradiction, we conduct a comprehensive examination of the existing PWA techniques in the literature (Table 1). Our analysis reveals that a significant portion of the prior papers contain one or more of four common pitfalls: (a) Data Leakage: where data samples from the same patient are present in both the train and test sets, (b) Overconstraining: where data far from normal range is discarded as outliers, which statistically simplifies the task, (c) Unreasonable Calibration: where the calibration method is not tested over longer (e.g.,> 1 day) time scales, and (d) Unrealistic Preprocessing: which filters a significant portion of the dataset terming it as noisy. We analyze these pitfalls in detail in our results section.

Our analysis reveal a somewhat surprising lack of improvement (modulo the pitfalls above) in PPG-based blood pressure prediction. This is in contrast to the substantive improvements in non-invasive prediction of other vitals such as heart-rate during this time. This raises the question as to whether there is a limit/ceiling on the prediction accuracy. In order to answer this, we propose tools to examine whether an input sensor signal (x) (e.g., PPG) can be a good predictor of the output health label (y) (e.g., BP). For this, we want to evaluate whether an underlying function f exists, which captures the relationship between x and y, such that $y = f(x)$. We also want to measure the *conditioning* of this underlying function, and check whether it is well-conditioned or not? That is, whether small changes in x lead to small or large changes in y. It is important to ensure that (minor) noise in the sensor measurement (which is inevitable in a real-world setting) does not lead to significant error in the outputs. Our tool is based on information-theoretic notions of *mutual information* and *multi-valued mappings*. Using our proposed tool, we find that BP prediction using PPG has a high multi-valued mapping factor of 33.2% and low mutual information of 9.8%. In comparison, heart rate prediction using PPG, a well-established task, has a very low multi-valued mapping factor of 0.75% and high mutual information of 87.7%. This confirms that estimating BP from PPG is a challenging and an ill-conditioned problem and a more principled approach is needed in the future for framing such health measure prediction tasks.

Results

In this section, we present a systematic review of prior work predicting BP via PPG PWA (Figure 1), followed by a principled analysis using our proposed tools.

Review of the results and limitations of prior work

To motivate our work, we analyzed recent research[21–23,27,29,34,48–51] that reported results predicting BP via PPG PWA (see Table 1). These works relied on the MIMIC[52] dataset (Appendix C.1) containing continuous PPG signals and the corresponding arterial BP values. They evaluated their performance against the AAMI[53] and/or BHS[54] standards (Appendix C.2). We found that they were prey to some common pitfalls, which resulted in misleading claims and over-optimistic results. For simplicity, we focus on the prediction of Systolic BP (SBP) rather than Diastolic BP (DBP), as SBP has a wider statistical range.

Before we begin, we should note that not all work (e.g.,[35–38,50]) followed the AAMI/BHS standards accurately. For example, some reported results on a test-set of fewer than 85 subjects. Moreover, although these works use the same MIMIC dataset, we found a lack of standardization in the train-test data splits and different BP ranges used for evaluation (due to differences in how the data were filtered) across the literature[27,29]. With the absence of official source code, it was difficult to reproduce prior results and compare different methods. Hence, we trained our own reference deep learning model (Figure 2), similar to the methods presented in prior research[27,34,49]. The *reference network* takes a three-channel input consisting of the original PPG waveform, along with its first and

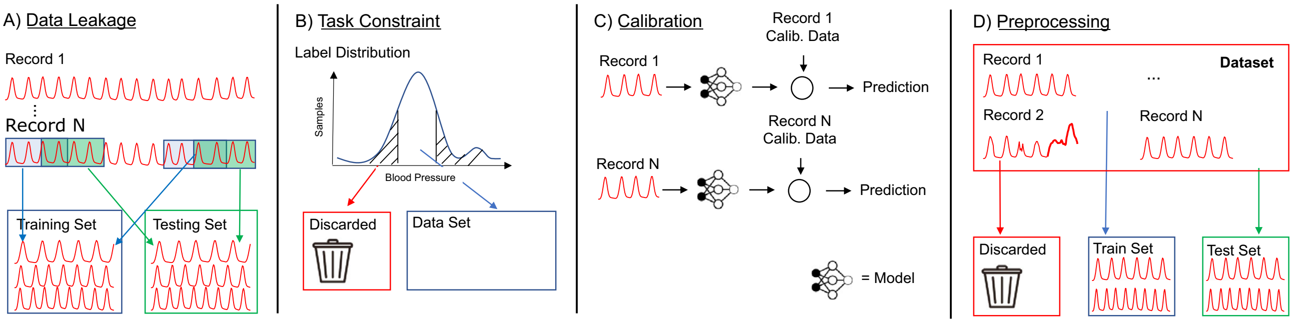


Figure 1. When designing end-to-end machine learning models researchers often use techniques such as: (**A**) providing the model with observations from similar patients, (**B**) constraining the task (e.g., limiting the distribution of labels), (**C**) calibrating models using data from a participant. When doing so it can often be difficult to identify how these steps impact the integrity of a model, or (**D**) preprocessing to filter out problematic samples (e.g., noisy inputs).

second derivatives, and outputs the predicted SBP value. The model consists of an eight-layer residual CNN[55] with 1D convolutions, and is trained using a mean squared error loss. We also explored 2D convolution based CNN models, such as DenseNet-161[28] and ResNet-101[55], taking spectrogram of the 1D PPG signal[27] and/or raw waveform as input. Among these, we found that the 1D CNN based architecture performed best.

Data leakage

The goal of any machine learning model is to generalize well to *test* data that will be seen in real-world settings[56]. Even with a large training set, it is very unlikely that identical samples to those seen in the training set will appear at test time, thus generalization is crucial. Unfortunately, good performance on a training dataset does not always translate to good performance on a test set, as models can *overfit*. This is especially true for modern deep neural networks, which are highly over-parameterized and can easily memorize the training data[57]. Thus, evaluating test performance accurately is an important step in understanding how a model will function in the real world. For this, the test data needs to be pristine, i.e., without any contamination from the training data. Unfortunately, contamination can and does happen in several ways.

We observed two types of overlap between training and testing splits (Figure 1A): *data-overlap* and *domain-overlap*.

Data-overlap corresponds to overlap of actual segments from a sample between the train-test sets. Domain-overlap is more subtle, where although there is no direct overlap of samples, leakage may occur due to similarities in train-test data. In our case, it corresponds to using different records from the *same* patient in both the test and train sets (Figure 3).

Here, we consider a particular example from the literature, PPG2ABP[21], where the authors propose a U-Net based architecture to predict the ABP (Arterial BP) waveform from PPG. They obtain impressive results with a bias of −1.19 mmHg and error standard deviation (SD) of 8.01 mmHg (Note, there is an error in the computation of standard deviation in the PPG2ABP[21] evaluation script. We report the corrected results here.) on the SBP prediction task (Table 2), which is close to the AAMI standard. However, while analyzing their source code, we found both data and domain overlaps.

Data-Overlap: The PPG2ABP[21] data processing pipeline divides each PPG *record* (∼6 mins long) into 10-second windows with an overlap of 5 seconds (URL: github.com/nibtehaz/PPG2ABP/blob/master/codes/data_processing.py) (Figure 3). Using overlapping windows helps, as it increases the size of the training data. However, the problem arises when these 10-second samples are randomly split into train and test sets. Since the overlapping windows are generated *before* the random train-test split, the train and test sets can have samples with the *same* overlapping regions (Figure 3). A deep learning model can memorize values based on these overlapping portions, leading to artificially high accuracy on the test set.

Domain-Overlap: Due to the physiological differences between individuals, person-dependent models often outperform person-independent models[58]. For example, for the BP prediction task, a model can learn the normal range of an individual's BP and leverage that to provide more accurate predictions. Since the knowledge of an individual's identity can impact a model's accuracy, it is important that the identity of the subject is not leaked (even implicitly) between test and train sets, especially while building person-independent models. Since the PPG signature has been shown to identify an individual[59], the presence of PPG signals from the same individual in both train and test data can thus leak identity. This turns out to be the case in the PPG2ABP work[21], as they randomly split PPG records into test and train sets, resulting in different windows from the *same* patient present in both test and train sets (Figure 3).

To quantitatively evaluate the impact of data leakage, we compare the performance of the PPG2ABP network on three splits (Figure 3) – (1) *No-overlap*: the dataset is partitioned at the patient level with an 80-20% train-test split, (2) *Domain-Overlap*: each patient has multiple records (∼6 mins long), and these records are randomly split 80–20% between the train-test set, i.e., records from the same patient can be present in both the training and test sets, and (3) *Data-Overlap*: We use the split provided by PPG2ABP[21] which divides the records into overlapping windows followed by an 80-20% train-test split. All splits consist of 10-second windows with an overlap of

Method	Dataset	Results (SBP)	Data-split	Over-constraining	Unrealistic Pre-Proc.	Calibration
BiGRU Attention[30]	MIMIC-II	MAE=2.58 SD=3.35	U	N SD=14.1	N ~10%	–
AdaBoost[31]	MIMIC-II	ME=0.09 MAE=8.22 SD=10.38	N D.O	Y	Y	–
ANN[32]	MIMIC-II	MAE=3.21 RMSE=4.23	U	Y	N ~75%	–
LSTM[33]	MIMIC-II	MAE=3.23 STD=4.75	U	U	Y	–
Ensemble CNN[34]	MIMIC-III	MAE=9.43	Y S.T	Y	N ~1.7%	–
ANN[35]	MIMIC	MAE=4.02 SD=2.79	U S.T	U	Y	–
Regression[36]	Custom Dataset	MAE=6.90 SD=9.00	Y S.T	Y	Y	–
SVR[37]	Queensland	ME=11.6 SD=8.20	Y S.T	Y	Y	–
Regression[38]	Custom Dataset	MAE=3.90 SD=5.37	Y S.T	N	Y	–
ANN[39]	MIMIC-II	ME=0.16 MAE=4.47 SD=6.85	N C.O	U	Y	–
SVR[40]	Custom Dataset	ME=5.10 SD=4.30	Y S.T	N SD=11.9	Y	–
SVR[41]	Queensland	MAE=4.76 SD=7.68	N D.O S.T	N	Y	–
Math Models[42]	Custom Dataset	MAE=7.66	Y S.T	N SD=12.5	Y	–
ANN[43]	MIMIC	MAE=3.80 SD=3.46	U S.T	N	N	–
Regression[44]	MIMIC	MAE=4.90 SD=6.59	N D.O S.T	Y	Y	N
LSTM-CNN[45]	MIMIC-II	ME=1.55 SD=5.41	U S.T	N	N ~15%	–
AdaBoost[46]	MIMIC-II	ME= -0.05 SD=8.90	N D.O	Y	N ~20%	–
U-Net[21]	MIMIC-II	ME=-1.58 SD=8.61	N C.O	Y	Y	–
CNN Siamese[27]	MIMIC-II	MAE=5.95 SD=6.90 [Calib]	Y	N	N ~5%	N
U-Net[29]	MIMIC-II	ME=4.30 SD=6.50	Y	N SD=13.5	Y	–
1-D CNN[47]	Custom Dataset	SD=11.4	N D.O S.T	N SD=16	Y	–
LSTM[48]	MIMIC-II	ME=4.05 SD=4.60	N D.O S.T	N	N ~50%	–

Table 1. The table summarizes the limitations of previous research and indicates whether the study exhibits specific pitfalls. The pitfalls are categorized into four categories: a) Data-split: Domain Overlap (denoted as D.O), Data Overlap (denoted as C.O), or Small test set (denoted as S.T). b) Over-constraining: SBP values and standard deviation (if provided) c) Unrealistic Pre-Processing: % of remaining dataset after pre-processing (if provided) d) Calibration: Correctly employed and justified for longer periods. The columns denote the presence or absence of each limitation, with Y (yes) indicating that the study has the limitation, N (no) indicating that it does not have that limitation, U (unknown) indicating that there is not enough information available, and "-" indicating that the research is not applicable to the pitfall. For additional information, please see Section Review of the results and limitations of prior work.

5-seconds to maintain consistency with the split proposed in PPG2ABP. Table 2 shows the performance of the PPG2ABP network over the three splits. Domain-overlap significantly increases the accuracy of the PPG2ABP network from a standard deviation of 23.1 to 16.2 mmHg; Data-Overlap further improves the standard deviation to 8.01 mmHg. This analysis clearly shows that leakages, however subtle, can lead to seemingly high but artificial improvements. Note that for all analysis in the rest of this paper, we use the *No-Overlap* split.

Overconstraining the task

Health-related data typically have non-uniform Gaussian distributions, with the highest data density near the "normal" (or healthy) range, and falling exponentially as we move away from the normal. We observe a similar trend for BP data in both the Aurora-BP[60] (Appendix C.1) and MIMIC datasets (see Figure 4). While points far

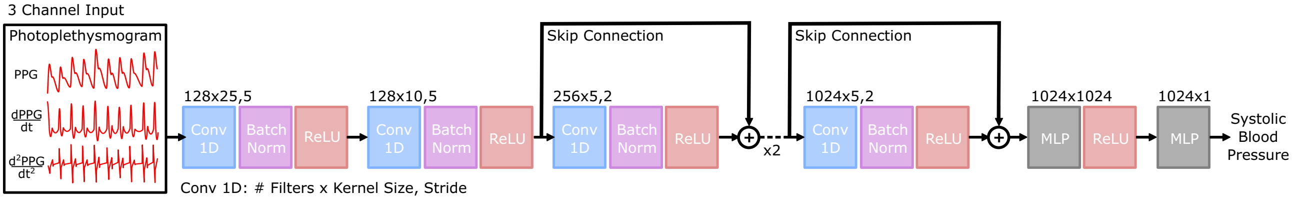
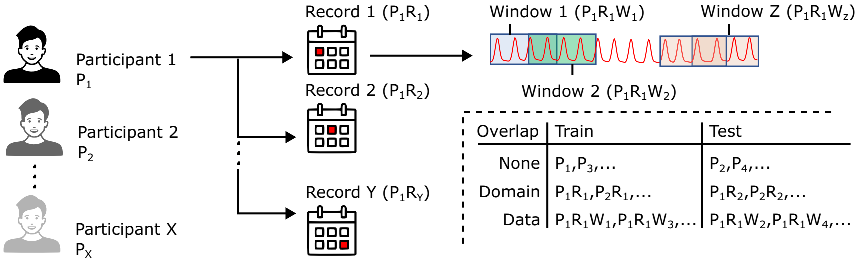


Figure 2. Our reference network, is used to evaluate the impact on performance due to the issues mentioned in Section "Review of the Results and Limitations of Prior Work". The network has 28M trainable parameters, takes a 3-channel input (PPG, VPG, APG), and outputs the SBP prediction. The model is optimized using a mean squared error loss.



Overlap	Train	Test
None	P_1,P_3,...	P_2,P_4,...
Domain	P_1R_1,P_2R_1,...	P_1R_2,P_2R_2,...
Data	$P_1R_1W_1$,$P_1R_1W_3$,...	$P_1R_1W_2$,$P_1R_1W_4$,...

Figure 3. Every participant (P) has multiple data records (R), and each record is divided into multiple overlapping windows (W). Each window forms a data *sample*. In No-Overlap, the train and test data are split at the participant level, while in Domain-Overlap, the split happens at the record level, and in Data-Overlap, the split happens at the window level.

	PPG2ABP [21]	
Data Split	**Bias (mmHg)**	**SD (mmHg)**
No-Overlap	1.11	23.1
Domain-Overlap	5.12	16.2
Data-Overlap	−1.19	8.01

Table 2. Performance of PPG2ABP[21] on different test-train splits with varying degrees of dataset overlap. Even subtle leakages can result in large (but artificial) accuracy improvements.

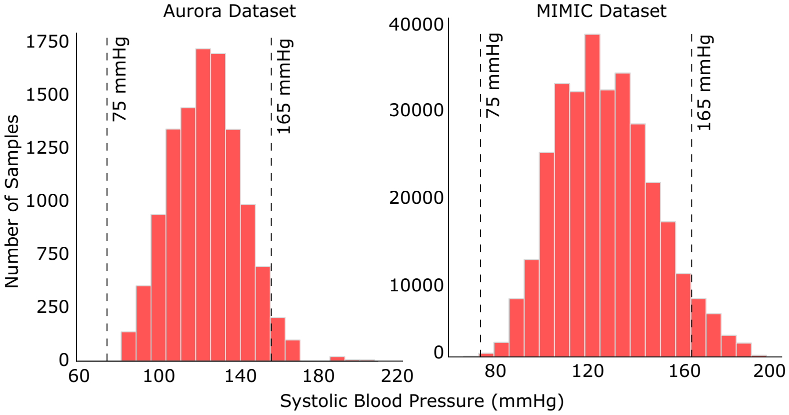


Figure 4. The distribution of systolic BP values in the: (left) Aurora-BP dataset and (right) MIMIC dataset. In the MIMIC dataset, the SBP values lie in the range 65–200 mmHg, however prior works ignore samples with SBP values outside the range of 75–165 mmHg.

from normal are rare, they are often crucial events (abnormally low or high BP) indicating serious health issues requiring medical attention.

However, we found that researchers often discard so-called "outliers"[22,27,29] (Figure 1B), arguing that such samples are unlikely or have occurred due to noise in the data collection process. For example, the MIMIC dataset has SBP values ranging between 65 and 200 mmHg (75-220 mmHg in Aurora-BP), but Schlesinger et al.[27] ignored samples outside the range of 75–165 mmHg, referring to the discarded values as "improbable". Similarly, Cao et al.[22] and Hill et al.[29] use a constrained range of 75–150 mmHg, while according to the British Hypertension Society literature, 140–159 mmHg is Grade-1 (mild) hypertension, 160–179 mmHg is Grade-2 (moderate) hypertension, and ≥180 mmHg is Grade-3 (severe) hypertension[54].

Constraining the data range has two problems. First, it leads to an incomplete evaluation, as the model is neither trained nor tested on samples from the discarded ranges. Second, since the statistical range of the output is reduced, this makes the prediction task artificially "easier" (i.e., a lower error can be achieved more easily), which may result in promising but misleading results. To quantitatively study the impact of constraining data ranges, we conducted an experiment using our reference network with different filtering of the data range. Table 3 shows the performance of our network when trained with three different SBP ranges: 65–200, 75–165 and 75–150 mmHg. Even small restrictions in the output range can lead to a significant (perceived) improvement in accuracy, e.g., reducing the SBP upper limit from 165 to 150 mmHg results in an ∼11.4% improvement in the standard deviation. This can be explained as samples at the extremes often result in the highest prediction errors (as models tend to predict closer to the mean of the distribution making predictions on samples with very high or low ground-truth BP values the most inaccurate).

The exclusion of samples with SBP measurements outside the range ≥165 mmHg and ≤75 mmHg during the training of machine learning models may result in overlooking crucial physiological features, potentially concealing serious health conditions and introducing bias into the model. This practice not only limits the scope of the developed models but also hinders conclusions about their generalizability and real-world applicability, as they become less representative of the diverse patient populations they are intended to serve.

Unreasonable calibration

The relationships between health measures (e.g., PPG and BP) are often person dependent. For example, blood pressure (bp) is dependent on the patient's heart rate (hr), blood viscosity ($visc$), stiffness of blood vessels ($stif$), etc., i.e., $bp = f(hr, visc, stif, \ldots)$. While the PPG signal might capture heart rate well, it may not be able to capture viscosity- and stiffness-related information. To solve this problem, it is common to propose the use of a calibration step, wherein a few PPG samples from each patient along with gold-standard BP values are used to calibrate the function f for that patient (Figure 1C). The model then learns a calibrated function, $\hat{f}$, for a specific patient, i.e., $bp = \hat{f}(hr)$, where the patient-specific parameters ($visc$, $stif$, ...) are folded into $\hat{f}$.

The literature does not offer a universally effective calibration strategy. Cao et al.'s[22] method needs to be calibrated every time before a BP prediction to find the optimal fit on the wrist for the watch, while Schlesinger et al.'s[27] model needs to be calibrated once to find the offset value between the model and the true prediction. As blood pressure may not change drastically within minutes (at rest) and significant trends might be observed only over the course of a few months owing to lifestyle changes or the influence of medication[61], it becomes important to pay attention to questions such as: What is the frequency of re-calibration? Is the calibration approach prone to changes in other environmental factors? We believe that the calibration approaches reported in prior work risk over-fitting by memorizing patient-level local temporal characteristics, and that evaluation is incomplete given that they do not evaluate BP prediction over longer time scales.

To understand the influence of calibration, we evaluate the prediction performance under different calibration strategies. *Naïve Calibration* simply predicts a constant calibrated value for the entire record. The constant value is computed as the mean of the ground truth values of the first three windows of a record. *Offset Calibration* uses our reference network, but adds an offset to the predicted value. The offset is computed in the calibration step as the difference between the predicted and ground truth BP of the test record's first window. We found the Naïve Calibration to perform very well (Table 4), with a standard deviation of 8.61 mmHg, close to the AAMI standard. However, predicting a constant BP value for a patient is clearly incorrect. This inconsistency underscores problems with the evaluation methodology. Since typical records in MIMIC have short time intervals (average length = 6 minutes) compared to the time scales at which BP changes, predicting a constant value gives deceivingly good accuracy. An appropriate evaluation of calibration methods should consider time scales spanning the intended re-calibration duration. For example, if re-calibration is planned every six months, the method should be evaluated with patients tracked over at least a six month time period. To demonstrate that calibration systems can quickly deteriorate over time, we analyzed the performance of Offset Calibration as the

SBP Range (mmHg)	Bias (mmHg)	SD (mmHg)
65–200	−3.45	15.8
75–165	−4.59	14.0
75–150	−4.42	12.4

Table 3. Performance of the reference network on different SBP ranges on the MIMIC dataset. Constraining the data range can result in significant (but artificial) accuracy improvements.

Method	Bias (mmHg)	SD (mmHg)
Naïve calibration	0.79	8.61
Offset calibration	0.38	9.82
No calibration	0.28	10.9

Table 4. Performance of different calibration-methods on the MIMIC dataset. The incorrect Naïve calibration methods perform very well, underscoring problems with the evaluation methodology.

time from the calibration window increases. Although the method performs well for the first few days, the error rates increase dramatically after that (Figure 5A).

Unrealistic preprocessing or filtering

The MIMIC dataset comprises ICU-patients data, with artifacts due to patient movement, sensor degradation, transmission errors from bedside monitors, and human errors in post-processing data alignment. The impact of these artifacts is visible in both the PPG and ABP waveforms as missing data, noisy data, and sudden changes in amplitude and frequency (Figure 6). To clean the signal, researchers[27,29] have used band-pass filters to remove noise in the high frequency (≥16 Hz) and low frequency (≤0.5 Hz) ranges, followed by auto-correlation to filter signals that are not strongly correlated with themselves. The auto-correlation step removes samples with uneven amplitude and/or frequency. After cleaning the MIMIC dataset (Figure 1D), Schlesinger et al.[27] used less than 5% of the total data for training their neural network, while Hill et al.[29] and Slapnicar et al.[34] used less than 10% of the total MIMIC data. This suggests that "clean" data is rare. Although filtering datasets to remove some noise is often an essential step to train a machine learning model[56], excessive filtering of data can result in overfitting. Models trained on such clean data might achieve high performance on a clean test set; however, they might fail in practice, as it is difficult to obtain such clean signals in a real-world scenario.

To understand the impact of filtering on a dataset, we measure the performance of our reference network at different auto-correlation thresholds. Figure 5(B) plots the performance of our reference network in predicting SBP

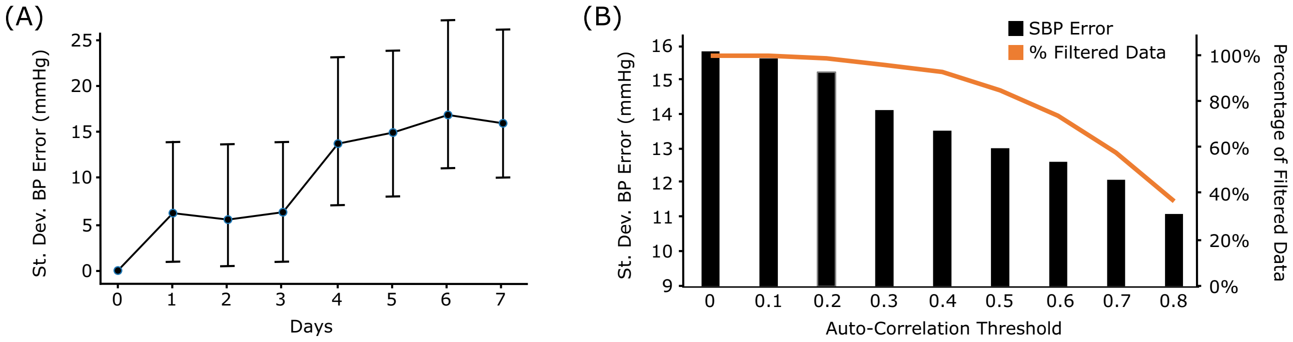


Figure 5. (**A**) The offset calibration method's performance falls off quickly after the first few days. (**B**) Performance of our reference network with different auto-correlation thresholds on the MIMIC dataset.

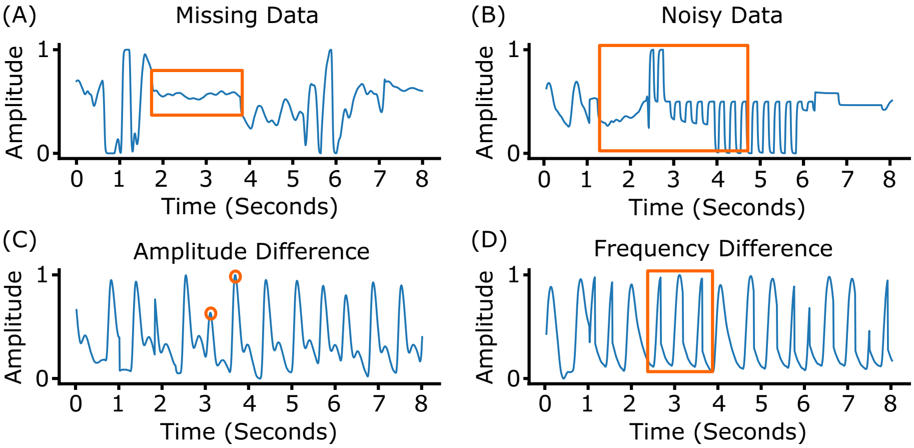


Figure 6. Examples of poor-quality photoplethysmography signals from the MIMIC dataset.

and the percentage of filtered data for each auto-correlation threshold. The performance of the network improves by 29.7% and the dataset size decreases by 63%, as we increase the auto-correlation threshold from 0 to 0.8.

Our proposed principled approach

We propose and utilize two tools—based on multi-valued mappings and on mutual information (Appendix B)—to estimate if the input signal is a good predictor of the output. Using our proposed tools we performed a principled analysis to study the relationship between PPG and BP. For comparison, we also used our tools on heart rate (HR) and reflected wave arrival time (RWAT) estimation for which it is known that the PPG signal is a strong predictor.

Checking for Multi-Valued Mappings: We use Algorithm 1 to find multi-valued mappings corresponding to data samples that are close in the input space but distant in the output space. As discussed in Section B.1, to compute the distance between two PPG inputs, we first align them using cross-correlation, followed by computing their Euclidean distance. We divide the dataset records into non-overlapping two-second windows and treat them as individual inputs. We set an input distance threshold of 1.0, which corresponds to a per-time sample threshold of $4e - 3$ (each 2s PPG window had 250 samples). For the output, we set thresholds of 8 mmHg, 8 bps, and 0.02s for the BP, HR and RWAT prediction tasks, respectively. We found very few multi-valued mappings for the HR and RWAT tasks, while a large number of mappings for the SBP task (Table 5). In the MIMIC dataset, for 33.2% of the 2-second windows, we found another window for the same patient who was close in the input PPG space but had a significantly different SBP output. When limiting the search to different patients, for 15.0% of the windows we could still find such matches. This implies that the task of predicting BP from PPG is ill-conditioned. Figure 7 shows examples of such multi-valued mappings, with highly similar input PPG waveforms but significantly different output arterial BP waveforms. In comparison, for the HR and RWAT tasks, the number of such matches is much smaller at 0.02% and 0.08% intra-patient, respectively, suggesting much better conditioning.

In the process of filtering multi-valued mappings, it is essential to consider the specificity of sensors and the methodologies employed in preprocessing the input data. Our analysis focuses on intra-patient and inter-patient multi-valued mappings within specific datasets, namely MIMIC and AURORA, rather than across different datasets. This approach ensures that our findings are not confounded by variations in sensor quality or the nuances of measurement techniques. Additionally, it enables us to apply preprocessing steps that preserve amplitude information.

Evaluating Mutual Information: To estimate mutual information (MI) between the PPG signal and the target output (BP/HR/RWAT), we use the K-nearest neighbours based approach proposed by Kraskov et al.[62]. We leverage dimensionality reduction to make MI estimation tractable, using handcrafted and auto-encoder learned feature representations. We report the mutual information of the input features and target variable, as well as the entropy of the target variable. Note that the target variable's entropy is the maximum achievable mutual information. Thus, the ratio of MI and target variable entropy represents the target information fraction encoded by the input, which we call *Info-Fraction*. We found *Info-Fraction* to be a more intuitive measure than the absolute MI values, and use it to compare the predictive power of PPG across the different tasks.

Handcrafted Features: As suggested by Takazawa[63] and Elgendi et al.[64], we calculate handcrafted features (see Table 6) from the PPG waveform (Figure 8). Due to the absence of a time-aligned ECG waveform in the MIMIC dataset, we extracted the relevant handcrafted features only from the PPG waveform. Table 7 presents the MI of these individual features with respect to the BP prediction task for both the MIMIC and Aurora-BP datasets, along with the MI when all these features are combined and regarded as a single multi-dimensional input. We found that even the combined features set encode a small fraction of the total target entropy. For example, in the MIMIC dataset, the combined features' *Info-Fraction* is just 9.5%, while heart rate itself contributes an *Info-Fraction* of 4.1%. Similar observations hold true for the Aurora-BP dataset. This hints that the PPG signal does not have enough information to predict BP in this dataset, and moreover the prediction is highly dependent on the heart rate.

For the Aurora-BP dataset we have the demographic data (age, weight, height) of the subjects, as well as time-aligned PPG and ECG waveforms. This allows us to calculate additional features, e.g., radial Pulse Arrival Time (rPAT) and other derived features[60]. Prior work[7] has used PAT to estimate blood pressure. Moreover, the Aurora-BP dataset has multiple readings for each subject in different positions (e.g., sitting, at rest, and supinated) which helps us add delta features reflecting the difference between features in the two conditions. Despite this, we found the entropy results for the Aurora-BP dataset to be similar to the MIMIC dataset, with the handcrafted features able to capture only 9.8% of the entropy of blood pressure (Table 8). On the other hand, for the HR and RWAT prediction tasks, the handcrafted features captured 87.7% and 64.6% entropy, respectively (ground truth for HR is derived from the ECG sensor data and RWAT from the tonometric sensor data). This

	MIMIC		Aurora-BP	
Task	Intra-patient	Inter-patient	Intra-patient	Inter-patient
SBP	33.2%	15.0%	13.9%	16.2%
HR	0.75%	2.10%	0.02%	0.89%
RWAT	–	–	0.08%	4.78%

Table 5. Multi-valued mapping matches for the BP, HR and RWAT prediction tasks. For the BP task, there was a high match rate for both within the same patient records and across patients, suggesting an ill-conditioned problem. For the HR and RWAT tasks, the matches were much lower. Ground truth for RWAT is only available for the Aurora-BP dataset.

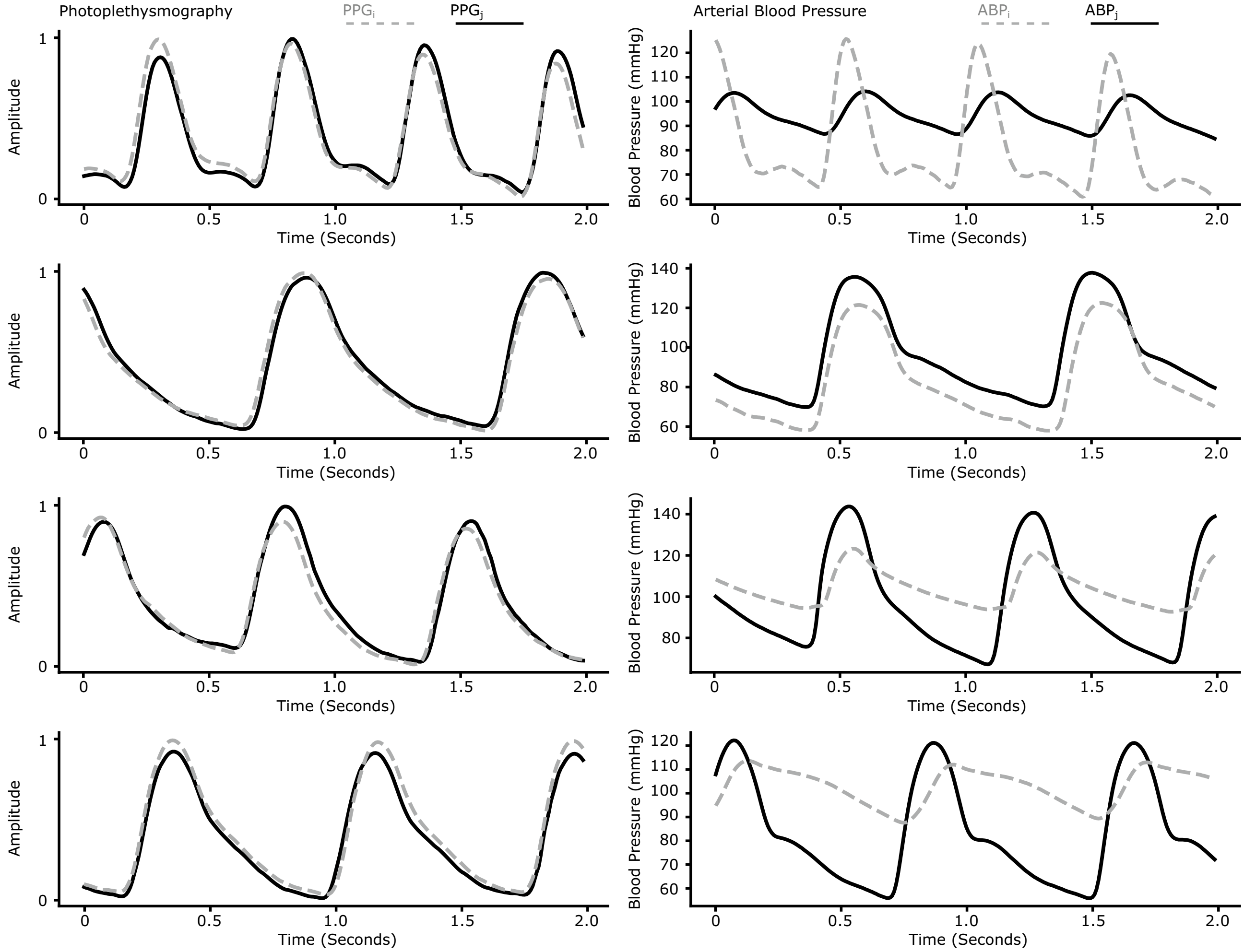


Figure 7. Multi-valued mappings. Examples of PPG waveforms (PPG_i and PPG_j) that are very similar and have corresponding arterial blood pressure waveforms (ABP_i and ABP_j) that are quite different. This highlights the existence of similar features that map to different targets, which makes the task of blood pressure prediction via PPG pulse wave analysis ill-conditioned.

Feature	Description
Heart Rate (HR)	Measurement of the number of pulsations of the heart in a minute. Calculated as the inverse of median time between each heart beat. The PPG signal was used for MIMIC (because time alignment with the ECG signal was not precise), the ECG signal was used for Aurora-BP.
Heart Rate Variability (HRV)	Measurement of the variation in time between each heart beat. Calculated as the mean of standard deviations of normal-normal (NN) intervals (SDNN).
Quality	Measurement of the quality of the PPG signal. A heuristic based algorithm that takes the signal-to-noise ratio, artifacts, consistency between the pulses in a window into consideration and computes a normalized score between 0 and 1.
$\frac{dp}{dt}$	Measurement of the mean systolic rise times normalized with respect to the duration of each beat in the PPG signal.
rPAT	Measurement of the delay between the R-peak in the ECG signal and systolic peak of the PPG signal. This can only be computed for Aurora-BP due to imprecise synchronization in MIMIC.
Inv. PAT	1/rPAT.
Δ Feature	Measured as the difference between the calculated value and baseline value. A baseline value of each feature for all patients is computed in Aurora-BP (not available for MIMIC).
std.Feature	Measures the fluctuation of a feature across a fixed time period.

Table 6. Descriptions of the handcrafted features used for the Mutual Information analyses.

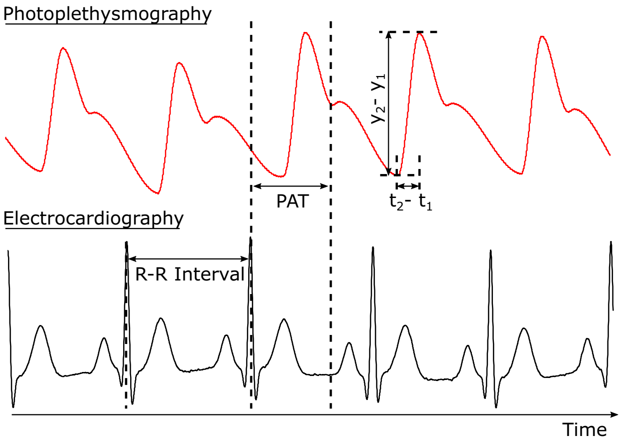


Figure 8. A visual description of the hand-crafted features calculated from the PPG and ECG waveforms. The systolic ramp time ($\frac{dp}{dt}$) is defined as $\frac{y_2 - y_1}{t_2 - t_1}$.

Optical Features	Mutual Information (bits)	
	MIMIC	Aurora-BP
HR	0.120	0.103
HRV	0.070	0.054
Quality	0.070	0.112
$\frac{dp}{dt}$	0.013	0.064
Combined	0.280	0.240
Entropy	2.930	3.680
Info-fraction (Combined)	9.5%	6.5%
Info-fraction (HR)	4.1%	2.8%

Table 7. Mutual Information of PPG optical features in the BP prediction task. Even all features combined have a small *Info-Fraction*, and most of that is captured by the heart rate feature alone.

further strengthens our finding that the PPG signal even with additional information from the ECG waveform has limited information to predict BP.

Auto-encoder Features: As an alternative to handcrafted features, we train an auto-encoder on the raw PPG waveform to obtain a set of low dimensional features. We use a five-layer perceptron (MLP) auto-encoder with ReLU activation and a bottleneck layer of 20 neurons. The model was trained with the Adam optimizer (learning rate of 0.001) and a mean-squared error loss (with a stopping point when the loss saturated at <0.1). Training time on a single NVIDIA P100 was under an hour. Table 9 shows the MI of the combined bottleneck features with respect to the BP, HR and RWAT prediction tasks. Although the auto-encoder features are more comprehensive and have higher MI compared to the hand-crafted features, the *Info-Fraction* for BP prediction (12.9% for MIMIC and 8.7% for Aurora-BP) is still much lower compared to that for HR (92.2% for MIMIC and 93.1% for Aurora-BP) and RWAT (70.1% for Aurora-BP) prediction tasks.

There are two possible implications of these findings. First, it may suggest that PPG signals lacks adequate information for accurate BP prediction. Alternativelty, it could imply a limitation in the current sensor technology, making sensors susceptible to confounding factors like external noise and environmental variations, thereby hindering the accuracy of BP prediction.

Conclusion

Our results reveal that BP prediction via pulse wave analysis of the PPG signal is still an unsolved task and far from the acceptable AAMI and BHS standards. By performing a systematic review and accompanying experiments we found several issues being overlooked in the prior work that have led to seemingly over-optimistic results. These pitfalls can be categorized into data splits that leak information from test samples into the training set, heavy constraints on the task that remove challenging samples and reduce the range of target values substantially, calibration methods that seem to be practically problematic, and unreasonable preprocessing that filters the data to an unrealistic extent such that any noise is unacceptable. These pitfalls simplify the machine learning task, creating a deceptive perception of ease in model training, which results in inflated performance. Ultimately, this translates to models that overfit the training data, hindering their ability to generalize effectively and handle real-world data variations.

	Mutual Information		
Feature	SBP (bits)	HR (bits)	RWAT (bits)
Age	0.026	0.015	0.000
Weight	0.024	0.024	0.000
Height	0.007	0.000	0.000
HR	0.130	2.000	0.658
std HR	0.009	1.230	0.509
rPAT	0.200	0.220	0.116
HRV	0.170	0.295	0.131
Inv. PAT	0.070	0.132	0.065
Quality	0.100	0.080	0.000
$\frac{dp}{dt}$	0.016	0.476	0.584
std $\frac{dp}{dt}$	0.009	0.232	0.252
Δ rPAT	0.160	0.165	0.074
Δ Inv. PAT	0.060	0.072	0.063
$\Delta\frac{dp}{dt}$	0.014	0.131	0.139
Δ HRV	0.170	0.130	0.016
Δ HR	0.025	0.242	0.234
Δ Quality	0.070	0.060	0.001
Combined	0.364	3.240	1.650
Entropy	3.680	3.650	2.540
Info-Fraction (Combined)	9.89%	88.8%	65.0%

Table 8. Mutual Information of patient demographic data, PPG optical features and features derived using ECG, for the Aurora-BP dataset[60]. While all features combined have an Info-Fraction of just 9.8% for the SBP prediction task, they encode much more information for the HR prediction (87.7%) and RWAT prediction (64.6%) tasks.

	MIMIC			Aurora-BP		
Task	MI	Entropy	*Info-Fraction*	MI	Entropy	*Info-Fraction*
SBP	0.38	2.93	12.9%	0.32	3.68	8.70%
HR	2.60	2.82	92.2%	3.40	3.65	93.1%
RWAT	–	–	–	1.78	2.54	70.1%

Table 9. Mutual information of auto-encoder features. The same trend of Table 8 holds. While the SBP task has low Info-Fraction, the features encode much more information for the HR and RWAT tasks. Ground truth for RWAT is only available for the Aurora-BP dataset[60].

While research on non-invasive approaches to estimate health vitals such as heart rate and blood oxygen saturation has made tremendous progress, enabling these technologies to become ubiquitous in the last decade, progress in non-invasive cuffless BP estimation has been slow despite witnessing similar research interest. This has prompted us to question whether the problem itself is ill-conditioned and if the PPG signal contains enough information to predict BP in the first place. In order to answer these questions, we have proposed a set of tools based on multi-valued mapping and mutual information to check if an input signal is a good predictor of the desired output. The multi-valued mapping checker allows us to find samples close in input space but far in output space. We found many such samples in both the MIMIC and Aurora-BP datasets. Searching for multi-valued mappings was trivial once appropriate distance metric and thresholds were defined, qualitative and quantitative results show that almost identical PPG waveforms can have very different BP waveforms. Next, we looked at the entropy of the features by computing mutual information. MI was extremely low for both hand-crafted and learned auto-encoder features. In comparison, heart rate and RWAT prediction tasks from PPG PWA have much lower multi-valued mapping factors and much higher mutual information indicating that the task is relatively well conditioned compared to PPG PWA to BP. We believe that these tools are relevant for feasibilty analysis in similar tasks involving wearable data, such as predicting stress levels from PPG[65–67] and estimating blood glucose levels from PPG[68–70].

Our study does not aim to prove that blood pressure estimation from PPG PWA is impossible; however, it indicates that the task is very challenging, and evaluating performance fairly is non-trivial. To navigate this complexity, we present a set of tools that future research can leverage to avoid the pitfalls identified here. We hope our work can serve as a milestone and stimulate further discussion and exploration in the following areas: (1) Data Diversity: Collecting comprehensive datasets that represent subjects from diverse demographics and cardiovascular physiologies. (2) Multiple modalities: Exploring the integration of PPG with other physiological signals holds immense potential for enhancing prediction accuracy and providing a more holistic view of cardiovascular health. (3) Improved Sensors: Advancements in sensor technology are crucial to capture higher-fidelity PPG data with minimal external noise and environmental variables. We believe that focusing on these critical areas will lead to generalizable and scalable solutions, empowering a future where everyone can benefit from the accessibility and convenience of non-invasive cuffless BP estimation.

Data availibility

All the data used in this work is publicly available. The MIMIC[71](https://archive.physionet.org/physiobank/database/mimic2wdb/) and Aurora-BP[60] (https://github.com/microsoft/aurorabp-sample-data) datasets can be accessed by researchers after completing the necessary steps stated by the creators of those datasets.

A Related work

The gold-standard for blood pressure measurement, used in Intensive Care Units and Operating Theatres, requires an invasive procedure that involves inserting a cannula needle into an artery. The cannula needle is connected to a transducer that converts the pulse signal to the arterial pressure waveform, providing continuous pulse-level BP measurements. Such invasive measurement is not feasible outside of a hospital setting, therefore two alternative cuff-based non-invasive procedures—auscultatory and oscillometry methods—are widely used [72]. However, these methods do not provide continuous measurement, Hence researchers [7,73,74] have been actively working on developing novel methods to accurately estimate blood pressure in a non-invasive continuous manner. A majority of the proposed methods involve calculating the Pulse Transit Time (PTT) which is inversely correlated to BP. PTT is defined as the time taken by a pulse to travel between two arterial sites—one measured using PPG and the other captured from a different sensor. E.g., Ding et al.[75] captured ECG, He et al. [73] used Ballistocardiogram from the ear, Holz and Wang [74] collected accelerometer signals from the head, and Wang et al. [7] captured accelerometer signals using a smartphone pressed to the chest.

Considering the ease and accessibility of accurately measuring heart rate and heart rate variability via PPG captured from a smartphone or wearable, a natural extension is to attempt to calculate blood pressure solely by analyzing the PPG pulse wave. Recent works [22,27,29,48,50] have explored and published promising results for the BP prediction task from PPG pulse wave analysis. These proposed methods involve building data-driven regression models to learn meaningful features by leveraging the availability of large PPG-BP labelled datasets (MIMIC [52]). For example, Schlesinger et al. [27] predicted BP using Convolution Neural Networks (CNN) trained on a frequency domain representation of the PPG signal and used siamese logic to calibrate BP predictions at run-time, Tazarv and Levorato [50] used a Long Short-Term Memory (LSTM) network with the PPG waveform as input, and Slapnicar et al. [34] proposed an ensemble network of 1-D CNNs and LSTMs on the raw and first two derivatives of the PPG signal. Some recent works [21,29] have proposed an extension to prior work by predicting the full Arterial Blood Pressure (ABP) waveform from the PPG signal using U-Net based architectures.

B Methods

We propose two tools—based on multi-valued mappings and on mutual information—to estimate if the input to a model is a good predictor of the output.

B.1 Multi-valued mapping check

If the input sensor signal (x) is a good predictor of an output health labels (y), it means there exists a function f, such that $y = f(x)$. Moreover, the function f should be well-conditioned, i.e., small changes in x should not lead to large changes in y. This is important to ensure that small amounts of noise in the sensor measurement (which are bound to happen in a real-world setting) do not lead to significant errors in the output. To test whether a task is well-conditioned, we propose searching for multi-valued mappings using Algorithm 1. Our multi-valued mapping algorithm searches for samples that are close in the input space but distant in the output space. If the algorithm is able to find such mappings, it means that the function f either does not exist, or is at best ill-conditioned.

```
$multivalued\_mappings$ = { }
ppg[i], sbp[i] = ppg wave at index i, sbp value at index i
$N$ = size of ppg and sbp arrays
dist($x_1$, $x_2$) = function to calculate distance between ppg wave $x_1$ and $x_2$
$t_i$ = threshold for input space
$t_o$ = threshold for output space
for $i = 1, 2, \ldots, N$ do
    for $j = 1, 2, \ldots, N$ do
        if $dist$(ppg[i], ppg[j]) $\leq t_i$ and $abs$(sbp[i]-sbp[j]) $\geq t_o$ then
            $multivalued\_mappings$.add([i,j])
        end if
    end for
end for
```

Algorithm 1. Multi-valued Mapping Search

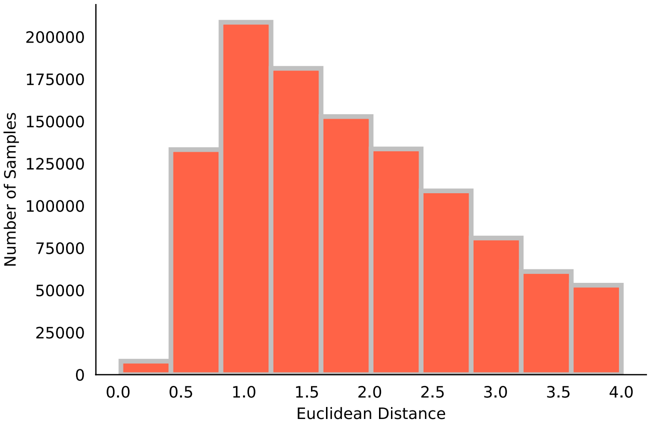


Figure 9. The distribution of Euclidean Distances between pairs of aligned consecutive PPG waves.

Algorithm 1 has two key components: a distance function for comparing the input samples and an optimal threshold value for filtering the multi-valued mappings.

Distance Function: Searching for multi-valued mappings in a dataset requires a metric to quantify the distance between the input samples. However, choosing the right distance function is not always obvious, and one needs to be careful about the implicit assumptions in any given metric. For example, cross-correlation, dynamic time warping (DTW) [76], and Euclidean distance are ways to measure the distance between two time-series/waveforms, and each has specific characteristics—cross-correlation is phase invariant, DTW is scale invariant in the time dimension, and Euclidean distance is translation invariant. For cross-correlation, a sliding window dot-product of the two input data series is computed to find the point where the similarity is maximized; DTW computes an optimal match by reducing the minimum-edit distance between the two series; Euclidean distance measures the similarity between the two data series using the L2 distance.

Ideally, the distance function should align well with the task requirements. Among the three distance functions, DTW makes the similarity metric invariant with respect to the time scale. However, it is known that BP has a direct dependency on heart rate, which in turn is determined by the periodicity of the PPG waves. Thus, the time scale invariance property of DTW will result in information loss for this task, making it a bad choice as a distance function. The Euclidean distance used in isolation is not a good choice either, as even the same PPG signals slightly shifted in time can result in a high Euclidean distance value. Since the relationship between PPG and BP should not change with small shifts of the PPG signal forward or backward in time, such a distance metric is not suitable. Therefore, cross-correlation is ideal to create an appropriate distance metric. Although the cross-correlation based distance metric worked well in our experiments, we found that aligning PPG signals via cross-correlation followed by computing the Euclidean distance between the aligned signals appeared logical. We used this distance measure for all our experiments.

Optimal Threshold: After choosing the appropriate distance function, we need to identify an optimal distance threshold, below which two signals can be considered “equal”. However, it is not straightforward to find such a threshold. If the threshold is very generous (i.e., high), we will end up selecting distant input signals as equal, and obtain misleading multi-valued mappings. On the other hand, if the threshold is too strict (i.e., low), we may not find any multi-valued mappings even for ill-conditioned functions, as the chances of two input signals being identical, especially in the presence of noise, are very small. To identify the optimal threshold for filtering multi-valued mappings, we calculate the Euclidean distance between two consecutive aligned PPG waves, each 2 seconds in duration. This interval was chosen because it represents an ideal time frame in which the signal remains consistent. Ideally, the difference between 2 consecutive PPG waves should account for an irreducible error, and this can be used as a threshold for filtering multi-valued mappings. Figure 9 illustrates the results of this analysis, which indicates that a majority of the PPG waves exhibit a Euclidean distance of ≤ 1, which led us to choose 1.0 as the threshold for our experiment.

Note that our multi-valued mapping check is a one-way method, i.e., if we are able to find multi-valued mappings, it implies an ill-conditioned *f*; however not finding multi-valued mappings does not guarantee existence of a well-defined *f*. This is because Algorithm 1 may fail to find signals close in the input space due to sparsity of the dataset. The mutual information check discussed next provides a complimentary method.

B.2 Mutual information check

Mutual Information (MI) is an information theoretical measure of the dependence between two random variables X and Y, defined as:

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X), \end{aligned} \tag{1}$$

where H is the Shannon entropy function ($H(X) = -\sum_i p(x_i) log(p(x_i))$). For continuous analog data, it is computed via limiting density of discrete points (LDDP)[77]. The marginal entropies $H(X)$ and $H(Y)$ represent the

Grade	Cumulative % of data below SBP error threshold		
	≤ 5 mmHg	≤ 10 mmHg	≤ 15 mmHg
A	60%	85%	95%
B	50%	75%	90%
C	40%	65%	85%

Table 10. Grading scale of test devices as per British Hypertension Society (BHS).

amount of information needed to describe the outcome of the random variable. This is same as the uncertainty of the random variable. $H(X|Y)$ and $H(Y|X)$ are conditional entropies, and denote the amount of information needed to describe the outcome of one random variable when the value of the other variable is known. This can also be thought of as the amount of uncertainty left in one random variable when the other is known. The mutual information I can be then interpreted as the amount of information (or reduction in uncertainty) that knowing one variable provides about the other. For example, $I(X; Y)$ is zero if X and Y are independent, while it is maximum when X is a deterministic function of Y or vice-versa.

Mutual information can be an effective measure in our case to evaluate whether the input signal (x) can be a good predictor of the output health label (y). However, since the computation of MI relies on estimation of probability density functions of the random variables, it is non-trivial to robustly estimate the MI for high dimensional data such as the time series PPG data. To overcome this curse of dimensionality, we recommend the following dimensionality reduction approaches before computing the MI.

Auto-Encoder. Since MI is invariant under smooth invertible transformations of the variables, we propose using an auto-encoder to aggressively reduce the input space dimensionality. We train an auto-encoder with the least number of bottleneck features needed to achieve a target mean-squared reconstruction loss of 0.1 on the normalized dataset. For the MIMIC and Aurora-BP dataset, we achieved this target with a bottleneck size of 20, at which the MI estimation worked robustly.

Hand-Crafted Features. As an alternate solution to using an auto-encoder, we can use hand-crafted features extracted from the input signal based on prior literature [63,64] and use these features for MI estimation. For example, in the task of BP prediction from PPG signal, common features include normalized systolic slope, heart rate, heart rate variability, etc. The MI estimation process helps us understand the importance of each of these features both collectively and independently. Note that in the case of hand-crafted features, there is always the concern of completeness (i.e., if the features extracted enough information from the input needed for the task), thus we recommend the auto-encoder approach whenever possible.

C Analysis details

C.1 Datasets

Our work builds on two datasets, the properties of which are critical to understand the results of our work.

MIMIC II: The MIMIC II dataset contains records of continuous high-resolution physiological waveforms of the patients in the ICU, such as ABP, PPG, and ECG sampled at 125Hz. The dataset consists of 67,830 records of varying duration from 30,000 patients [71]. For the purpose of our study, we perform our analysis on a pre-processed subset of the MIMIC II dataset, consisting of 12,000 records from 942 patients[52]. This subset is particularly useful for our analysis as it includes a sufficient number of patients for training and testing, compliant with AAMI standards, and has been commonly utilized in previous research(Table 1).

Aurora-BP: The Aurora-BP dataset [60] consists of 24,650 records from 483 subjects. Each subject has multiple records of varying duration, which were collected at rest or while performing activities such as exercise and brisk walk. The records are collected from multiple sensors/devices including optical PPG, EKG, tonometer, accelerometer, and cuff-based Blood Pressure.

C.2 Performance standards

To contextualize the performance of SBP (Systolic BP) prediction task, two benchmarks have been widely used: AAMI and BHS standards. The criteria of the AAMI (Association for the Advancement of Medical Instrumentation) standards[53] are that the test set should comprise of at least 85 subjects, with at least 10% of them having an SBP above 180 mmHg and at least 10% having an SBP under 100 mmHg. For a test device to be compliant with the AAMI standards, the SBP prediction must have a bias under 5 mmHg and error standard deviation (SD) under 8 mmHg on the test set. The BHS (British Hypertension Society) [54] standards criteria states that the test set should consist of at least 85 subjects and that the cohort should be representative of the target audience of the device. The performance of the test device is divided into grades (Table 10). Additionally, the test data should cover the overall pressure range, specifically in these three ranges: ≤ 130, 130–160, and ≥ 160 mmHg.

C.3 Other considerations

Dataset size: To understand the effect of data size on MI, and verify whether our dataset had enough samples to enable robust MI estimation, we conducted the following experiment. We took a randomly selected slice of the data (ranging from 0.1 to 100% data) and computed the combined MI over 20 runs (this technique is known as bootstrapping). We performed this analysis for both the MIMIC and Aurora-BP datasets. As shown in Figures 10(A) and (B), although the estimates at smaller dataset sizes resulted in high variation, the variation bounds are very tight at higher sizes. This imparts confidence that our MI estimates over the full datasets are robust. Interestingly, we also found that using a smaller dataset can result in higher estimates of the MI values.

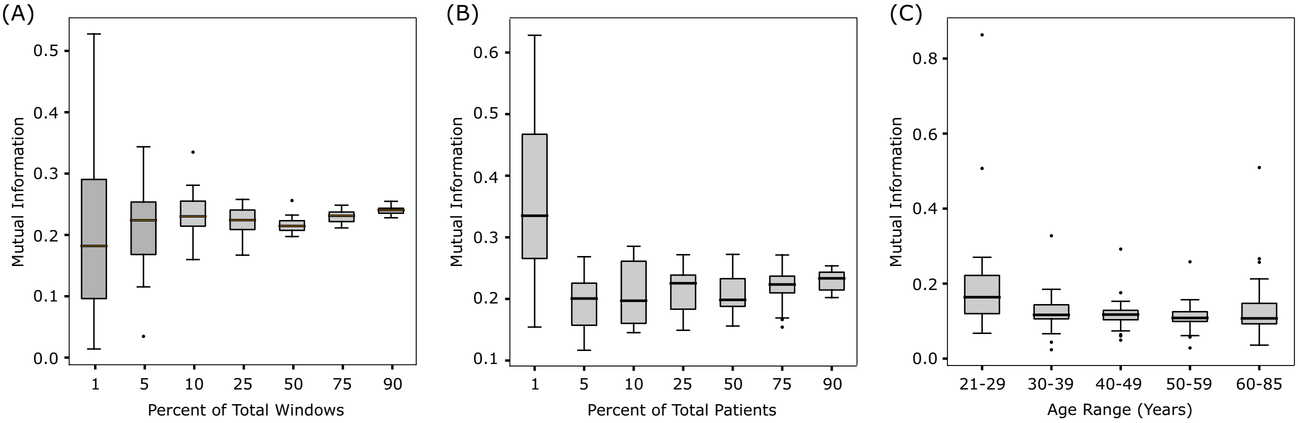


Figure 10. The effect of the number of (**A**) total windows (MIMIC dataset), (**B**) total patients (Aurora-BP dataset), and (**C**) age range (Aurora-BP dataset) on the mutual information between PPG PWA features and BP. We perform 20 runs with different random subsets of the data to plot the distributions. Optical PPG features similar to Table 7 were used for (**A**) and (**B**), while richer features (patient demographic data, PPG optical features and features derived using ECG, similar to Table 8) were used for (**C**). For each plot the corresponding features were combined and treated as a single multi-dimensional input for computing MI.

This may be explained by the fact that fewer multi-valued mappings might be observed in a smaller sample. Thus, having a small dataset might lead to an over optimistic perception of the relationship between input and output.

Participant's Demography: Apart from data size, we found that even demographic factors, such as age, impacted mutual information. Figure 10(C) shows the variation in combined MI with respect to age for the Aurora-BP dataset. In particular, we found that in the age group of 21-29 and 60-85 years, heart rate and weight were the most important features, which was not the case with the other age groups.

Received: 18 June 2024; Accepted: 29 July 2024
Published online: 07 August 2024

References

1. Bhat, K. S., Jain, M. & Kumar, N. Infrastructuring telehealth in (in)formal patient-doctor contexts. *Proc. ACM Hum.-Comput. Interact.***5**, https://doi.org/10.1145/3476064 (2021).
2. Haleem, A., Javaid, M., Singh, R. P. & Suman, R. Telemedicine for healthcare: Capabilities, features, barriers, and applications. *Sensors International* **2**, 100117. https://doi.org/10.1016/j.sintl.2021.100117 (2021).
3. Bestsennyy, O. Telehealth: A quarter-trillion-dollar post-covid-19 reality? (2021).
4. Patel, S. Take a pulse on health and wellness with your phone (2021).
5. Poh, M.-Z., McDuff, D. J. & Picard, R. W. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Opt. Express* **18**, 10762–10774 (2010).
6. Scully, C. G. *et al.* Physiological parameter monitoring from optical recordings with a mobile phone. *IEEE Trans. Biomed. Eng.* **59**, 303–306 (2011).
7. Wang, E. J. *et al. Seismo: Blood Pressure Monitoring Using Built-in Smartphone Accelerometer and Camera, 1–9* (Association for Computing Machinery, New York, NY, USA, 2018).
8. Wang, E. J. *et al.* Hemaapp: noninvasive blood screening of hemoglobin using smartphone cameras. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 593–604 (2016).
9. Gairola, S. *et al.* Smartkc: Smartphone-based corneal topographer for keratoconus detection. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.***5**, https://doi.org/10.1145/3494982 (2022).
10. Aggarwal, A. *et al. Towards automating retinoscopy for refractive error diagnosis* (Proc. ACM Interact. Mob, Wearable Ubiquitous Technol, 2022).
11. Liu, X. *et al.* Mobilephys: Personalized mobile camera-based contactless physiological sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, Volume Issue 1, March 2022, Article No.: 24*https://doi.org/10.1145/3517225 *(2022). arXiv:2201.04039.*
12. Liu, X., Fromm, J., Patel, S. N. & McDuff, D. Multi-task temporal shift attention networks for on-device contactless vitals measurement. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. & Lin, H. (eds.) *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual* (2020).
13. Fuchs, F. D. & Whelton, P. K. High blood pressure and cardiovascular disease. *Hypertension* **75**, 285–292. https://doi.org/10.1161/hypertensionaha.119.14240 (2020).
14. Sun, D. *et al.* Type 2 diabetes and hypertension. *Circ. Res.* **124**, 930–937 (2019).
15. Mukkamala, R. *et al.* Toward ubiquitous blood pressure monitoring via pulse transit time: Theory and practice. *IEEE Trans. Biomed. Eng.* **62**, 1879–1901 (2015).
16. Buxi, D., Redouté, J.-M. & Yuce, M. R. A survey on signals and systems in ambulatory blood pressure monitoring using pulse transit time. *Physiol. Meas.* **36**, R1-26 (2015).
17. Sharma, M. *et al.* Cuff-less and continuous blood pressure monitoring: A methodological review. *Technologies Basel* **5**, 21 (2017).
18. Bramwell, J. C. & Hill, A. V. The velocity of pulse wave in man. *Proceedings of the Royal Society of London. Series B, Containing Papers of a Biological Character***93**, 298–306, https://doi.org/10.1098/rspb.1922.0022 (1922).
19. Wesseling, K., Jansen, J., Settels, J. & Schreuder, J. Computation of aortic flow from pressure in humans using a nonlinear, three-element model. *J. Appl. Physiol.* **74**, 2566–2573 (1993).
20. Mukkamala, R., Stergiou, G. S. & Avolio, A. P. Cuffless blood pressure measurement. *Annu. Rev. Biomed. Eng.* **24**, 203–230 (2022).
21. Ibtehaz, N. & Rahman, M. S. Ppg2abp: Translating photoplethysmogram (ppg) signals to arterial blood pressure (abp) waveforms using fully convolutional neural networks (2020). arXiv:2005.01669.

22. Cao, Y., Chen, H., Li, F. & Wang, Y. *Crisp-BP: Continuous Wrist PPG-Based Blood Pressure Measurement, 378–391* (Association for Computing Machinery, New York, NY, USA, 2021).
23. Meneguitti Dias, f. *et al.* A machine learning approach to predict arterial blood pressure from photoplethysmography signal. In *Computing in Cardiology Conference (CinC)* (Computing in Cardiology, 2022).
24. Han, M. *et al.* Feasibility and measurement stability of smartwatch-based cuffless blood pressure monitoring: A real-world prospective observational study. *Hypertens. Res.* **46**, 922–931 (2023).
25. Groppelli, A. *et al.* Feasibility of blood pressure measurement with a wearable (watch-type) monitor during impending syncopal episodes. *J. Am. Heart Assoc.*https://doi.org/10.1161/jaha.122.026420 *(2022).*
26. Mukkamala, R. *et al.* Evaluation of the accuracy of cuffless blood pressure measurement devices: challenges and proposals. *Hypertension* **78**, 1161–1167 (2021).
27. Schlesinger, O., Vigderhouse, N., Eytan, D. & Moshe, Y. Blood pressure estimation from ppg signals using convolutional neural networks and siamese network. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1135–1139, https://doi.org/10.1109/ICASSP40776.2020.9053446 (2020).
28. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2261–2269, https://doi.org/10.1109/CVPR.2017.243 (2017).
29. Hill, B. L. *et al.* Imputation of the continuous arterial line blood pressure waveform from non-invasive measurements using deep learning. *Sci. Rep.* **11**, 15755. https://doi.org/10.1038/s41598-021-94913-y (2021).
30. El-Hajj, C. & Kyriacou, P. Deep learning models for cuffless blood pressure monitoring from PPG signals using attention mechanism. *Biomed. Signal Process. Control* **65**, 102301. https://doi.org/10.1016/j.bspc.2020.102301 (2021).
31. Hasanzadeh, N., Ahmadi, M. M. & Mohammadzade, H. Blood pressure estimation using photoplethysmogram signal and its morphological features. *IEEE Sens. J.* **20**, 4300–4310. https://doi.org/10.1109/jsen.2019.2961411 (2020).
32. 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. https://doi.org/10.3390/s20195668 (2020).
33. Hajj, C. E. & Kyriacou, P. A. Cuffless and continuous blood pressure estimation from PPG signals using recurrent neural networks. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, https://doi.org/10.1109/embc44109.2020.9175699 (IEEE, 2020).
34. Slapničar, G., Mlakar, N. & Luštrek, M. Blood pressure estimation from photoplethysmogram using a spectro-temporal deep neural network. *Sensors (Basel)* **19**, 3420 (2019).
35. Wang, L., Zhou, W., Xing, Y. & Zhou, X. A novel neural network model for blood pressure estimation using photoplethesmography without electrocardiogram. *J. Healthc. Eng.* **1–9**, 2018. https://doi.org/10.1155/2018/7804243 (2018).
36. Dey, J., Gaurav, A. & Tiwari, V. N. InstaBP: Cuff-less blood pressure monitoring on smartphone using single PPG sensor. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, https://doi.org/10.1109/embc.2018.8513189 (IEEE, 2018).
37. Zhang, Y. & Feng, Z. A SVM method for continuous blood pressure estimation from a PPG signal. In *Proceedings of the 9th International Conference on Machine Learning and Computing*, https://doi.org/10.1145/3055635.3056634 (ACM, 2017).
38. Jain, M., Deb, S. & Subramanyam, A. V. Face video based touchless blood pressure and heart rate estimation. In *2016 IEEE 18th International Workshop on Multimedia Signal Processing (MMSP)*, https://doi.org/10.1109/mmsp.2016.7813389 (IEEE, 2016).
39. Gaurav, A., Maheedhar, M., Tiwari, V. N. & Narayanan, R. Cuff-less PPG based continuous blood pressure monitoring — a smartphone based approach. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, https://doi.org/10.1109/embc.2016.7590775 (IEEE, 2016).
40. Gao, S. C., Wittek, P., Zhao, L. & Jiang, W. J. Data-driven estimation of blood pressure using photoplethysmographic signals. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, https://doi.org/10.1109/embc.2016.7590814 (IEEE, 2016).
41. Duan, K., Qian, Z., Atef, M. & Wang, G. A feature exploration methodology for learning based cuffless blood pressure measurement using photoplethysmography. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, https://doi.org/10.1109/embc.2016.7592189 (IEEE, 2016).
42. Suzuki, A. Inverse-model-based cuffless blood pressure estimation using a single photoplethysmography sensor. *Proc. Inst. Mech. Eng. [H]* **229**, 499–505. https://doi.org/10.1177/0954411915587957 (2015).
43. Kurylyak, Y., Lamonaca, F. & Grimaldi, D. A neural network-based method for continuous blood pressure estimation from a PPG signal. In *2013 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*, https://doi.org/10.1109/i2mtc.2013.6555424 (IEEE, 2013).
44. Slapnicar, G., Lustrek, M. & Marinko, M. Continuous blood pressure estimation from PPG signal. *Informatica (Slovenia)***42** (2018).
45. Panwar, M., Gautam, A., Biswas, D. & Acharyya, A. PP-net: A deep learning framework for PPG-based blood pressure and heart rate estimation. *IEEE Sens. J.* **20**, 10000–10011. https://doi.org/10.1109/jsen.2020.2990864 (2020).
46. Mousavi, S. S. *et al.* Blood pressure estimation from appropriate and inappropriate PPG signals using a whole-based method. *Biomed. Signal Process. Control* **47**, 196–206. https://doi.org/10.1016/j.bspc.2018.08.022 (2019).
47. Shimazaki, S., Kawanaka, H., Ishikawa, H., Inoue, K. & Oguri, K. Cuffless blood pressure estimation from only the waveform of photoplethysmography using cnn. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, https://doi.org/10.1109/embc.2019.8856706 (IEEE, 2019).
48. Harfiya, L. N., Chang, C.-C. & Li, Y.-H. Continuous blood pressure estimation using exclusively photopletysmography by LSTM-based signal-to-signal translation. *Sensors*https://doi.org/10.3390/s21092952 *(2021).*
49. Shimazaki, S., Kawanaka, H., Ishikawa, H., Inoue, K. & Oguri, K. Cuffless blood pressure estimation from only the waveform of photoplethysmography using CNN. *Annu Int Conf IEEE Eng Med Biol Soc* **2019**, 5042–5045 (2019).
50. Tazarv, A. & Levorato, M. A deep learning approach to predict blood pressure from PPG signals. *CoRR***abs/2108.00099** (2021). arXiv:2108.00099.
51. Mahmud, S. *et al.* A shallow u-net architecture for reliably predicting blood pressure (bp) from photoplethysmogram (ppg) and electrocardiogram (ecg) signals (2021). arXiv:2111.08480.
52. Kachuee, M., Kiani, M. M., Mohammadzade, H. & Shabany, M. Cuff-less high-accuracy calibration-free blood pressure estimation using pulse transit time. In *2015 IEEE International Symposium on Circuits and Systems (ISCAS)*, 1006–1009, https://doi.org/10.1109/ISCAS.2015.7168806 (2015).
53. 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).
54. O'Brien, E. *et al.* European society of hypertension recommendations for conventional, ambulatory and home blood pressure measurement. *J. Hypertens.* **21**, 821–848 (2003).
55. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778, https://doi.org/10.1109/CVPR.2016.90 (2016).
56. Domingos, P. A few useful things to know about machine learning. *Commun. ACM* **55**, 78–87. https://doi.org/10.1145/2347736.2347755 (2012).
57. Zhang, C., Bengio, S., Hardt, M., Recht, B. & Vinyals, O. Understanding deep learning (still) requires rethinking generalization. *Commun. ACM* **64**, 107–115. https://doi.org/10.1145/3446776 (2021).

58. D'mello, S. K. & Kory, J. A review and meta-analysis of multimodal affect detection systems. *ACM computing surveys (CSUR)***47**, 1–36 (2015).
59. Karimian, N., Guo, Z., Tehranipoor, M. & Forte, D. Human recognition from photoplethysmography (ppg) based on non-fiducial features. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4636–4640, https://doi.org/10.1109/ICASSP.2017.7953035 (2017).
60. Mieloszyk, R. *et al.* A comparison of wearable tonometry, photoplethysmography, and electrocardiography for cuffless measurement of blood pressure in an ambulatory setting. *IEEE Journal of Biomedical and Health Informatics* (2022).
61. Hinderliter, A. L. *et al.* The long-term effects of lifestyle change on blood pressure: One-year follow-up of the ENCORE study. *Am. J. Hypertens.* **27**, 734–741. https://doi.org/10.1093/ajh/hpt183 (2013).
62. Kraskov, A., Stögbauer, H. & Grassberger, P. Estimating mutual information. *Phys. Rev. E*https://doi.org/10.1103/physreve.69.066138 *(2004).*
63. Takazawa, K. Clinical usefulness of the second derivative of a plethysmogram (acceleration plethysmogram). *J. Cardiol.* **23**, 207–217 (1993).
64. Elgendi, M. *et al.* The use of photoplethysmography for assessing hypertension. *NPJ Digit. Med.* **2**, 1–11 (2019).
65. Iqbal, T. *et al.* Stress monitoring using wearable sensors: A pilot study and stress-predict dataset. *Sensors (Basel)* **22**, 8135 (2022).
66. Celka, P., Charlton, P. H., Farukh, B., Chowienczyk, P. & Alastruey, J. Influence of mental stress on the pulse wave features of photoplethysmograms. *Healthc. Technol. Lett.* **7**, 7–12 (2020).
67. Elzeiny, S. & Qaraqe, M. Stress classification using photoplethysmogram-based spatial and frequency domain images. *Sensors (Basel)***20** (2020).
68. Zhang, G. *et al.* A noninvasive blood glucose monitoring system based on smartphone PPG signal processing and machine learning. *IEEE Trans. Industr. Inform.* **16**, 7209–7218 (2020).
69. Hossain, S. *et al.* Estimation of blood glucose from PPG signal using convolutional neural network. In *2019 IEEE International Conference on Biomedical Engineering, Computer and Information Technology for Health (BECITHCON)* (IEEE, 2019).
70. Bent, B. *et al.* Engineering digital biomarkers of interstitial glucose from noninvasive smartwatches. *NPJ Digit. Med.* **4**, 89 (2021).
71. Goldberger, A. L. *et al.* PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation* **101**, E215-20 (2000).
72. Bonnafoux, P. Auscultatory and oscillometric methods of ambulatory blood pressure monitoring, advantages and limits: a technical point of view. *Blood Press. Monit.* **1**, 181–185 (1996).
73. Da He, D., Winokur, E. S., Heldt, T. & Sodini, C. G. The ear as a location for wearable vital signs monitoring. In *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*, 6389–6392 (IEEE, 2010).
74. Holz, C. & Wang, E. J. Glabella: Continuously sensing blood pressure behavior using an unobtrusive wearable device. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* **1**, 1–23 (2017).
75. Ding, X.-R., Zhang, Y.-T., Liu, J., Dai, W.-X. & Tsang, H. K. Continuous cuffless blood pressure estimation using pulse transit time and photoplethysmogram intensity ratio. *IEEE Trans. Biomed. Eng.* **63**, 964–972 (2015).
76. Bellman, R. & Kalaba, R. On adaptive control processes. *IRE Trans. Autom. Control.* **4**, 1–9 (1959).
77. Jaynes, E. T. Information theory and statistical mechanics. *Phys. Rev.* **106**, 620–630. https://doi.org/10.1103/PhysRev.106.620 (1957).

Author contributions

S.M. performed analyses, designed experiments, and wrote the manuscript. N.K. designed the experiments and wrote the manuscript. M.J. designed the experiments and wrote the manuscript. D.M. designed the experiments and wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

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

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.