CBPE: Causal-aware Blood Pressure Estimation with Multimodal Data

Hoyoon Byun
hoyun.byun@yonsei.ac.kr
Yonsei University
Seoul, Korea

Sumin Park
2019122009@yonsei.ac.kr
Yonsei University
Seoul, Korea

Jinho Kang
bubble3jh@uos.ac.kr
University of Seoul
Seoul, Korea

Hyeonjeong Lee
ehj1224@etri.re.kr
Medical IT Convergence Laboratory, Electronics and Telecommunications Research Institute
Daegu, Korea

Minseong Kim
kms5905@etri.re.kr
Medical IT Convergence Laboratory, Electronics and Telecommunications Research Institute
Daegu, Korea

Doosik Kim
doosik@etri.re.kr
Medical IT Convergence Laboratory, Electronics and Telecommunications Research Institute
Daegu, Korea

Kwang-Yong Kim
kimky@etri.re.kr
Medical IT Convergence Laboratory, Electronics and Telecommunications Research Institute
Daegu, Korea

Kyungwoo Song*
kyungwoo.song@yonsei.ac.kr
Yonsei University
Seoul, Korea

Abstract

Photoplethysmography (PPG)-based blood pressure (BP) estimation remains challenging due to inherent signal variability and the indirect nature of the measurement. However, although BP prediction is a high-stakes domain where accuracy is crucial, existing studies often overlook multimodal data that is commonly available in real-world healthcare scenarios, thereby limiting estimation accuracy. To address this gap, we introduce the Signal-driven and Demographic-Clinical Attribute dataset for BP Estimation (SABP), which integrates diverse modalities, including physiological signals, structured clinical attributes, and disease-related medical text, to comprehensively capture causally linked factors related to BP. The multimodal nature of SABP facilitates a deeper understanding of complex interactions between BP and various clinical attributes across different data modalities. However, modeling these interactions can be challenging due to unobserved variables or hidden confounders that potentially bias the underlying interaction of BP and PPG waveforms. To mitigate this issue, we propose the Causal-aware Blood Pressure Estimator (CBPE). CBPE leverages a Variational Autoencoder (VAE) to infer latent representations of hidden confounders, thereby improving the accuracy and robustness of BP estimation by effectively reducing hidden confounding effects. To complement model evaluation, we introduce the Group Loss Consistency (GLC) metric, which jointly assesses predictive performance and fairness across different BP groups. Unlike traditional error metrics, GLC explicitly penalizes performance disparities, encouraging reliable estimation across patient categories. Experimental results show that CBPE achieves competitive BP prediction accuracy while demonstrating superior robustness to subgroup bias and performance degradation under challenging conditions. Reproducible materials are available at https://github.com/MLAI-Yonsei/CBPE

CCS Concepts

• **Applied computing** → **Health care information systems**; • **Computing methodologies** → *Artificial intelligence.*

Keywords

Blood pressure estimation, Multimodal dataset, Causal relationship

ACM Reference Format:
Hoyoon Byun, Sumin Park, Jinho Kang, Hyeonjeong Lee, Minseong Kim, Doosik Kim, Kwang-Yong Kim, and Kyungwoo Song. 2026. CBPE: Causal-aware Blood Pressure Estimation with Multimodal Data. In *The 41st ACM/SIGAPP Symposium on Applied Computing (SAC '26), March 23–27, 2026, Thessaloniki, Greece.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3748522.3780020

1 Introduction

Hypertension and hypotension pose serious health risks, emphasizing the need for regular blood pressure (BP) monitoring for early detection and intervention [1]. Photoplethysmography (PPG) is a convenient and non-invasive method for indirectly measuring BP, but it struggles to deliver accurate measurements of systolic BP (SBP) and diastolic BP (DBP). As a result, improving BP prediction precision from PPG signals has become an active area of research [2, 3]. Relying solely on PPG signals is insufficient for accurate BP estimation, as various physiological factors can affect both PPG and BP. For instance, conditions like cardiac arrhythmia or respiratory

*Corresponding author

SAC '26, Thessaloniki, Greece

ACM ISBN 979-8-4007-2294-3/2026/03
https://doi.org/10.1145/3748522.3780020

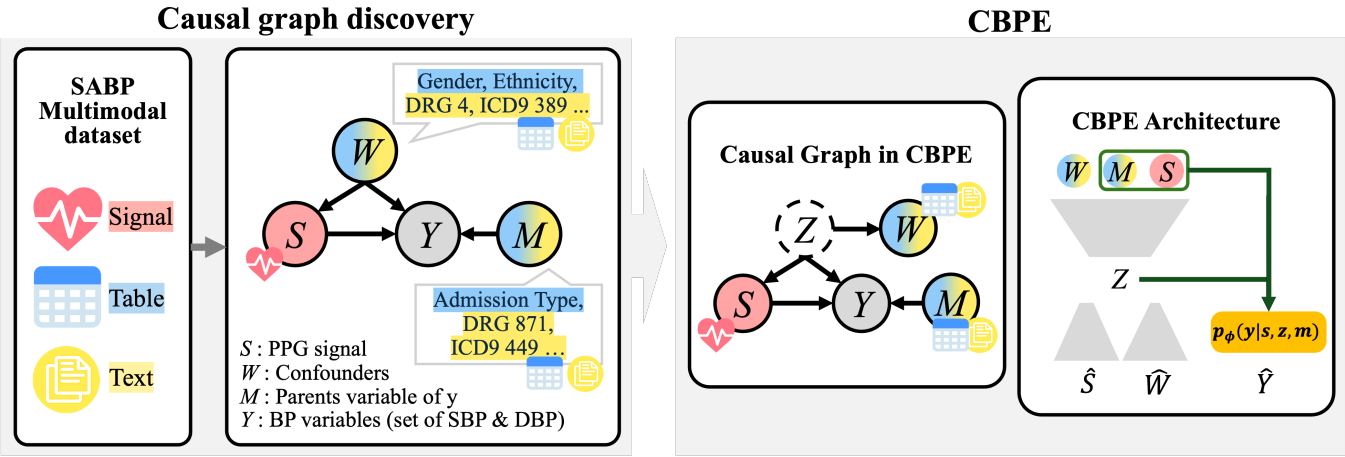

Figure 1: Overview of CBPE: We first discover a causal graph from the SABP multimodal dataset, identifying observed factors related to blood pressure (BP). To account for hidden confounders, CBPE augments the graph with a latent variable and estimates BP using a multimodal variational autoencoder that integrates signal, tabular, and text representations. Text modality refers to the DRG/ICD-9 code description, reflecting disease and clinical status information. For each patient, DRG or ICD-9 codes related to confounders and parents of y are merged into a single textual description, which is embedded using a text embedding model and concatenated with the other modalities to form the model input.

failure may act as confounders, distorting their relationship. Confounders that induce spurious correlations are a critical reason why estimators tend to be biased toward major groups or features, and why they often exhibit poor generalization under distribution shifts or in out-of-distribution settings [4, 5]. Our experimental results show that baselines that do not consider causal features yield less consistent performance across blood pressure groups. To address this, we first advance BP estimation by leveraging features that directly exert a causal effect on it. Building upon this, we further tackle the challenge that not all confounders are observable in practice, as prediction bias from unmeasured variables may still persist.

Accordingly, our second objective is to improve the precision of BP prediction by inferring hidden confounders from observable variables. For instance, the degree of stress serves as a factor influencing BP variations and can act as a confounder for both PPG and BP measurements. Furthermore, stress levels are one of the potential causes of cardiac arrhythmia and respiratory failure. As such, information related to these diseases can be utilized as proxies to estimate a hidden confounder, such as stress levels.

Existing benchmark datasets provide preprocessed PPG signals but lack rich clinical attributes [6, 7, 8, 9], whereas clinical databases such as MIMIC-III [10] offer detailed patient information but lack processed PPG data, limiting their use for BP modeling. To bridge this gap, we construct the Signal-driven and Demographic-Clinical Attribute dataset for BP Estimation (SABP) by integrating MIMIC-III with PulseDB [9]. SABP combines PPG signals with diverse clinical attributes relevant to BP estimation, enabling causal analysis of dependencies among BP, PPG, and patient attributes. However, simply incorporating causally related attributes yields limited performance gains.

Consequently, we hypothesize that hidden confounders influencing BP also bias BP estimators. To address this, we propose the novel framework, Causal-aware Blood Pressure Estimator (CBPE), which integrates PPG signals, clinical attributes, and a latent representation of hidden confounders inferred via a VAE [11]. The inferred latents are fused with PPG and BP-causal attributes to improve estimation. Extensive experiments show that CBPE achieves higher accuracy and robustness to data imbalance while maintaining consistent performance across BP groups—outperforming conventional methods in both stability and precision. Figure 1 illustrates the framework, including confounder inference and final BP prediction.

Conventional metrics such as MAE and RMSE measure overall accuracy but overlook disparities across BP groups. To ensure fair and consistent evaluation, we introduce the Group Loss Consistency (GLC) metric, which jointly captures accuracy and fairness by penalizing group-level performance imbalance.

2 Related works

2.1 Blood Pressure Estimation via Photoplethysmography

Recent studies have highlighted the strong link between PPG signals and blood pressure (BP) estimation, adopting various signal-based modeling approaches [12, 13]. Common methods treat PPG as a time-series input and employ forecasting or deep feature extractors such as TimeMixer [14], ResNet1D, MLP-BP [3], and SpectroResNet [15], as reported in benchmark analyses [16]. Few-shot transfer strategies, such as GloGen's dual-prompt framework [2], have been proposed to improve generalization. Meanwhile, multimodal approaches combining PPG with additional variables leverage inductive biases between physiological and demographic factors [17,

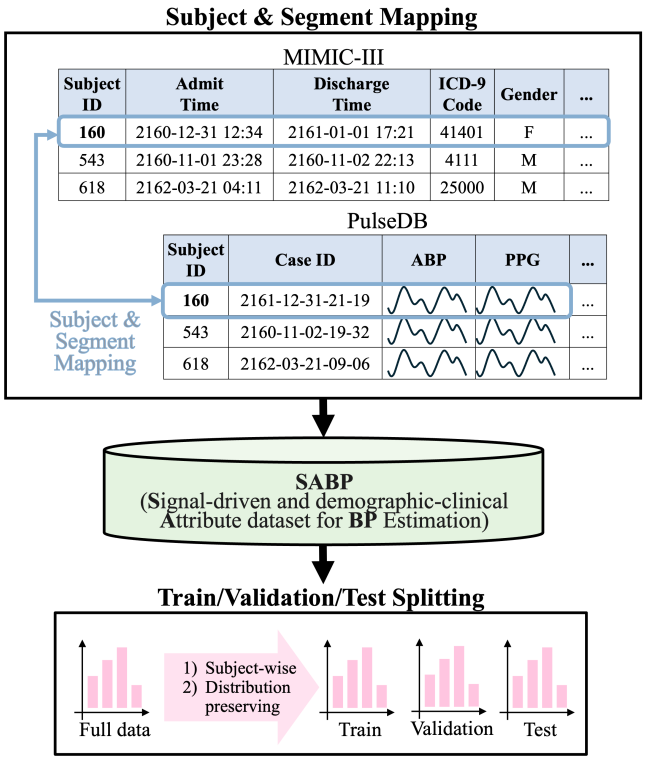

Figure 2: Construction and splitting of the SABP dataset. The dataset is built by integrating MIMIC-III clinical records with PulseDB physiological signals and is split into two stages to group by subject while preserving balanced blood pressure categories for generalizable evaluation.

18], with HGCTNet [18] fusing PPG, ECG, and pulse pressure waveforms to derive pulse transit time features. In addition, some studies have explored causal graph discovery for BP estimation by analyzing PPG signals [19]. However, these approaches focus on extracting features from PPG and inferring their causal relationships with BP, without considering various attributes of the patient.

2.2 Causal graph discovery

Causality represents the relationship in which one variable affects another [20, 21, 22], and is often modeled using causal graphs, specifically Directed Acyclic Graphs (DAGs) that characterize the causal dependencies between variables. Causal graph discovery aims to learn such structures from observed data and can be broadly categorized into constraint-based [23, 24] and score-based methods [25, 26, 27, 28].

Score-based approaches formulate discovery as an optimization problem, identifying the graph structure that maximizes a predefined scoring metric. Recent works integrate deep learning to improve scalability and performance. In our study, we adopt DAGMA [25] that enables efficient causal discovery via well-defined gradient updates. DAGMA demonstrates strong inference accuracy and computational efficiency, especially in large-scale graph settings.

Table 1: Summary of SABP Dataset Characteristics and Blood Pressure Statistics

Processed Amount	Validation Strategy
Subject: 1,572 Segment: 9,976 Duration: ~10 hours Seg./Sub.: ~16 *Sampling:* 125 Hz / 10 s	Subject-Wise and Distribution-Preserving Split ***Data Dist. (SBP/DBP, mmHg)***: Min.: 30.06 / 21.27 Max.: 211.14 / 176.57 Mean: 121.07 / 60.76 SD: 22.35 / 12.85
Data Attributes	
Signal-derived: PPG, ABP, SBP, DBP	***Clinical/Demographic***: Admission Location, Admission Type, Current Service, DRG Code, Ethnicity, Insurance, Religion, Sensitivity, Age, Gender, ICD-9

2.3 Causal representation learning

Research on causal effect estimation in deep learning has progressed rapidly [29, 30]. Early models such as TARNet [31] and Dragonnet [32] assumed that observed covariates X fully explain confounding in predicting Y. Subsequent approaches, starting with CEVAE [33], modeled X as proxies for latent confounders Z and applied VAE-based inference [11] to estimate Z, enabling scalable causal estimation in healthcare [34, 35]. Building on CEVAE, we focus on predictive modeling rather than effect estimation, targeting causal parents of Y, reducing spurious latents from over-assumed confounding, and extending the framework to regression over signals and tabular attributes. In parallel, causal perspectives are increasingly used to improve deep model generalization [36, 37, 38], though most works explore invariance or feature shifts without explicitly modeling causal relations among attributes. Causal approaches to multimodal regression that combine signal, tabular, and text data remain scarce. We introduce a multimodal latent variable method that integrates these modalities to improve regression accuracy.

3 Benchmark data construction

3.1 Data preprocessing.

We construct the Signal-driven and Demographic-Clinical Attribute dataset for BP Estimation (SABP) by integrating ICU clinical records from MIMIC-III [10] with curated physiological signals from PulseDB [9], which are derived from MIMIC-III and VitalDB [39]. After quality-filtering procedures that ensure reliable waveform inputs, high-quality physiological signals are aligned with contemporaneous patient attributes to form synchronized multimodal samples for BP estimation. The pipeline is shown in Figure 2.

From the comprehensive clinical records of MIMIC-III, we extracted 16 candidate attributes relevant to BP estimation, retaining 11 after excluding those with over 80% missingness. PulseDB includes “Subject ID” and "Case ID" fields, which map to the patient identifier and timestamp information in MIMIC-III, respectively.

This enables linking each PPG segment to the patient's clinical attributes at the time of acquisition. The resulting dataset is presented in Table 1, consistent with the format of [16].

3.2 Subject-wise & distribution preserving split.

We partition the full dataset into training, validation, and test sets following two principles: (1) ensuring that all records from a single subject are contained within the same fold to prevent information leakage, and (2) preserving the distributions of SBP and DBP values across folds to account for their inherent skewness and ensure adequate representation of all BP categories. The splitting process is consistent with prior work on BP benchmark datasets [16].

To implement these principles, each segment is first assigned to one of 16 BP classes by discretizing SBP and DBP into four levels each. Each subject is then labeled with a *Dominant BP Class*, defined as the most frequent class among their segments. Before splitting, we apply noise reduction to enhance data quality.

The splitting follows a two-stage process. First, we perform stratified subject-level partitioning based on the *Dominant BP Class*, giving priority to subjects from underrepresented classes to ensure adequate representation of rare BP classes. Second, we minimize class distribution discrepancies across folds by evaluating weighted differences (inverse-squared class counts) across 10,000 random seeds and selecting the optimal one. This yields subject-disjoint folds that closely preserve the original BP distributions. Subsequently, we downsample to 10,000 segments using stratified sampling to maintain class balance, aligning the dataset size with other BP datasets, such as Sensors [8], BCG [7], and PPGBP [6].

3.3 Multimodal causal graph discovery.

We apply the DAGMA algorithm [25] to discover causal structures among variables while optimizing a score function under acyclicity constraints. Inputs are built by reducing high-dimensional signals to two dimensions with t-SNE and aggregating them via a variance-weighted average. Categorical variables such as DRG and ICD-9 codes are filtered to retain the most frequent entries covering 50% of cumulative counts, resulting in 54 DRG and 123 ICD-9 codes, which are one-hot encoded. The final dataset includes 1,915 patients and 236 features, serving as the input for causal graph discovery.

4 Methodology

4.1 Causal-aware blood pressure estimation via representation learning of hidden confounder

We categorize the variables related to BP in the inferred causal graph into four types: (1) Parents that directly cause BP, (2) Confounders that affect both PPG and BP, (3) Indirect causal variables acting through intermediates, and (4) Non-causal variables. We focus on BP's parents and confounders. Figure 3a shows a simplified graph from observed data linking BP with its direct causes and with PPG. Let S, W, M, Y denote the PPG signal, confounders, parents of Y, and BP, respectively. Because hidden confounders of both S and Y are difficult to recover from observations, we treat observed variables as proxies and model hidden factors as VAE latents [33] (Figure 3b). To avoid introducing spurious associations by using all variables as proxies, we restrict proxies to variables with a causal effect on Y, which leads to more stable performance. We then train the VAE under the causal Markov assumption so that its factorization mirrors the underlying causal structure.

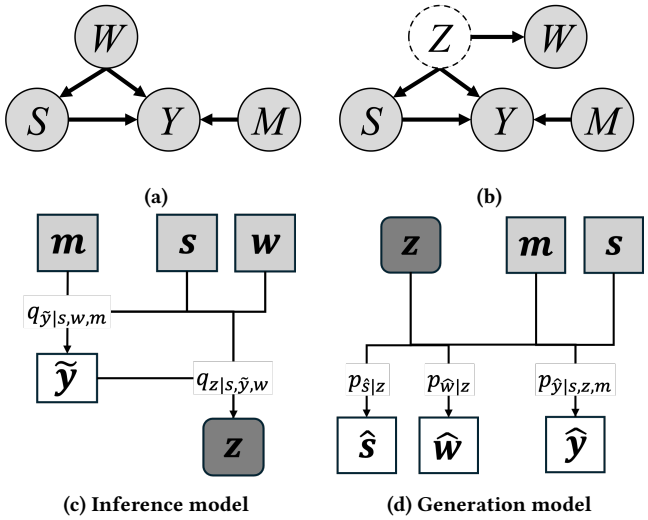

Figure 3: (a) Simplified causal graph without latent variable Z. (b) Simplified causal graph with the latent variable Z. (c) Inference model. $\tilde{y}$ is the output of the auxiliary BP estimator, modeled by $q_{\tilde{y}|s,w,m}$. $q_{z|s,\tilde{y},w}$ represents the posterior distribution over the latent variable z inferred from the sample. (d) Generation model. $p_{*|z}$ and $p_{*|s,z,m}$ represents the generation model.

Assumption 1 (Causal Markov Assumption). *In a causal Directed Acyclic Graph (DAG), each variable X_i is conditionally independent of any variables that do not belong to its parent or descendant set, given its parent variables. Formally, this is expressed as:*

$$X_i \perp\!\!\!\perp \mathbf{V} \setminus (Descendants(X_i) \cup Pa(X_i)) \mid Pa(X_i),$$

where $\mathbf{V}$ denotes the set of all variables in the DAG, $Pa(X_i)$ represents the set of parent variables of X_i.

We aim to maximize the Evidence Lower Bound (ELBO), employed here as a surrogate objective for approximating the true distribution of the observed data.

$$\text{ELBO} = \mathbb{E}_{q_\phi(z|s,y,w,m)}[\log p_\theta(s,y,w,m \mid z) + \log \frac{p_\theta(z)}{q_\phi(z \mid s,y,w,m)}].$$

We parameterize the model with generative parameters θ and variational parameters ϕ. $S \in \mathbb{R}^{d_s}$ is the PPG signal variable, $Y \in \mathbb{R}^2$ is the BP variable (a set of SBP and DBP), $W \in \mathbb{R}^{d_w}$ is a proxy variable (a set of confounders for PPG and SBP/DBP in attributes) for the latent variable $Z \in \mathbb{R}^{d_z}$, and $M \in \mathbb{R}^{d_m}$ represents the parents of Y (a set of attributes that affect only Y variables). In practice, S is taken as the output embedding from the penultimate layer of a Conv1D model, while structural/tabular attributes are one-hot encoded and passed through a lookup table whose embeddings are averaged per instance to form W and M. The text modality

is encoded using a text embedding model to obtain high-quality representations, followed by a linear projection to reduce the dimension, and we set the embedding dimensions to be equal, i.e., $\mathbb{R}^{d_m} = \mathbb{R}^{d_w} = \mathbb{R}^{d_z}$.

To learn the latent variable z, based on Assumption 1, we factorize the joint probability $p(s, y, w, m)$ by leveraging the identifiable causal relationships:

$$\begin{aligned}\text{ELBO} = \mathbb{E}_{q_\phi(z|s,y,w,m)}[&\log p_\theta(s \mid z) \\ &+ \log p_\theta(w \mid z) + \log p_\theta(y \mid s, z, m)] \\ &- D_{KL}(q_\phi(z \mid s, y, w, m) \parallel p_\theta(z)).\end{aligned} \tag{1}$$

By explicitly accounting for the causal relationships among variables, as reflected in the factorized ELBO (Equation 1), we maximize it using the observed data, where D_{KL} denotes the KL-divergence. This ensures that the latent variable Z effectively represents the hidden causal features of Y in the causal graph (Figure 3b).

Figure 3c and Figure 3d illustrate the inference and generation processes of CBPE. The final BP estimation value is given by $\hat{y} \sim p_\theta(y \mid s, z, m)$.

4.2 Auxiliary BP estimator

To maximize the ELBO, it is necessary to optimize the approximate posterior $q_{z|s,y,w,m}$. However, under evaluation, y cannot be given as a condition since it is the prediction target. Therefore, during model training, we incorporate an additional network to predict y, enabling its use as a conditioning variable in the posterior. The auxiliary loss for training this network is defined as follows:

$$\begin{aligned}\mathcal{L}_{Aux} &= -\mathbb{E}_{p(s,y,w,m)}[\log q_\psi(y \mid s, w, m)] \\ &= -\frac{1}{N}\sum_i p(s_i, y_i, w_i, m_i) \log q_\psi(y_i \mid s_i, w_i, m_i).\end{aligned}$$

N denotes the total count of training instances. ψ denotes the parameters of the auxiliary BP predictor. The final objective is calculated as the total of the ELBO and the auxiliary loss:

$$\mathcal{L}_{\text{CBPE}} = -\text{ELBO} + \mathcal{L}_{Aux}.$$

4.3 Theoretical analysis

Drawing on Pearl's do-operator [40], this section demonstrates that CBPE produces predictions immune to spurious correlations. By conditioning exclusively on the direct causal parents of the target variable Y, CBPE can be interpreted as an invariant predictor [41]. In graphical causal models, the do-operator formalizes an intervention: $do(X = x)$ is defined as severing all incoming edges to the node X and deterministically setting X to the fixed value x.

However, because the do-operator specifies an interventional distribution rather than a statistical estimand, it can be computed from observational data only after an identification step establishes an expression that depends solely on observable quantities.

Theorem 1. *(Identifiability of the conditional potential outcome) Let the treatment and parents features of y be denoted by s and m, respectively, and let the latent variable z represent the hidden confounder. Then the conditional distribution of the potential outcome is identifiable:*

$$Y^s \mid m \sim p\big(y \mid \text{do}(s), m\big) = \int_z p(y \mid s, z, m) p(z) dz$$

Definition 1 (Backdoor Criterion). *A set of variables z meets the backdoor criterion with respect to treatment s and outcome y if z blocks all backdoor paths from s to y, and z does not include any descendants of s.*

Proof. We begin by applying the backdoor criterion in Definition 1 to express the interventional distribution $p\big(y \mid \text{do}(s), m\big)$ in terms of observational data.

We identify a set of variables z (e.g., latent confounders) that satisfy the backdoor criterion with respect to s and y. Under this condition, intervening on s does not affect the distribution of z, i.e.,

$$p\big(z \mid \text{do}(s)\big) = p(z),$$

and the conditional distribution $p(y \mid s, z, m)$ reflects the causal influence of s upon y after adjusting for z.

Thus, the distribution under intervention can be rewritten using the backdoor adjustment formula:

$$p\big(y \mid \text{do}(s), m\big) = \int_z p\big(y \mid \text{do}(s), z, m\big)\, p\big(z \mid \text{do}(s), m\big)\, dz \tag{2}$$

$$= \int_z p(y \mid s, z, m)\, p(z)\, dz. \tag{3}$$

By Definition 1, conditioning on z blocks all spurious (backdoor) paths between s and y; hence, the second line holds, and the causal effect is identifiable from the distribution $p(s, y, m, z)$. □

Algorithm 1 CBPE Training and Inference Pipeline

Require: Dataset $\mathcal{D} = \{(s_i, w_i, m_i, y_i)\}_{i=1}^N$ ▷ s: PPG, w: confounder proxies, m: direct causes of y, y: SBP or DBP

Pre-step: Discover causal DAG with DAGMA [25] → identify M, W
Initialize VAE parameters θ, ϕ and auxiliary estimator ψ
for each epoch = 1 **to** T **do**
  for each mini-batch $b \subset \mathcal{B}$ **do**
    $\tilde{y} \leftarrow q_\psi(y \mid s, w, m)$ ▷ Auxiliary BP prediction
    $z \sim q_\phi(z \mid s, \tilde{y}, w, m)$ ▷ Posterior of latent confounder
    $\text{ELBO} \leftarrow \log p_\theta(s \mid z) + \log p_\theta(w \mid z) + \log p_\theta(y \mid s, z, m) - D_{KL}\big[q_\phi(z \mid s, y, w, m) \parallel p_\theta(z)\big]$
    $\mathcal{L}_{\text{Aux}} \leftarrow -\log q_\psi(y \mid s, w, m)$
    $\mathcal{L}_{\text{CBPE}} \leftarrow -\text{ELBO} + \mathcal{L}_{\text{Aux}}$
    Update $\{\theta, \phi, \psi\}$ with an optimizer to minimize $\mathcal{L}$
  end for
end for
function PredictBP(s, w, m) ▷ Inference time
  $\tilde{y} \leftarrow q_\psi(y \mid s, w, m)$
  $z \leftarrow \mathbb{E}_{q_\phi(z|s,\tilde{y},w,m)}[z]$
  $\hat{y} \leftarrow p_\theta(y \mid s, z, m)$
  return $\hat{y}$ ▷ Estimated SBP / DBP
end function

[0]Unlike CBPE, CEVAE treats all attributes as observed confounders without considering their causal relationships.

Table 2: All reported metrics represent the mean±standard deviation over 10 runs. We compare models using extracted PPG features (△) against multimodal approaches like CBPE that use raw signals with tabular data. CBPE outperforms across every metric, achieving the lowest error in both per-sample accuracy (S-loss) and, most notably, in cross-group consistency (G-loss). This significant lead in G-loss proves its ability to deliver robust and fair predictions by minimizing bias across patient groups.

Method	PPG Signal	Attributes	BP S-loss	SBP S-loss	DBP S-loss	BP G-loss	SBP G-loss	DBP G-loss	GLC
HGCTNet	△	O	$32.12_{\pm 3.66}$	$21.08_{\pm 2.93}$	$11.04_{\pm 1.28}$	$39.77_{\pm 4.08}$	$26.49_{\pm 3.11}$	$13.28_{\pm 1.43}$	76.58
TimeMixer	O	X	$27.16_{\pm 0.21}$	$17.60_{\pm 0.17}$	$9.56_{\pm 0.17}$	$36.03_{\pm 0.55}$	$24.00_{\pm 0.48}$	$12.03_{\pm 0.25}$	70.12
SpectroResNet	O	X	$30.72_{\pm 6.57}$	$19.53_{\pm 3.45}$	$11.19_{\pm 3.12}$	$38.65_{\pm 7.14}$	$25.06_{\pm 3.85}$	$13.60_{\pm 3.31}$	75.12
MLPBP	O	X	$28.77_{\pm 1.46}$	$18.50_{\pm 1.10}$	$10.27_{\pm 0.38}$	$37.07_{\pm 1.01}$	$24.28_{\pm 0.82}$	$12.79_{\pm 0.24}$	72.49
Conv1D	O	X	$30.70_{\pm 0.66}$	$19.76_{\pm 0.49}$	$10.94_{\pm 0.22}$	$37.39_{\pm 1.32}$	$24.38_{\pm 0.98}$	$13.01_{\pm 0.37}$	73.56
CEVAE	O	O	$30.23_{\pm 3.46}$	$18.49_{\pm 1.68}$	$11.73_{\pm 1.95}$	$38.34_{\pm 4.75}$	$24.40_{\pm 2.60}$	$13.94_{\pm 2.29}$	62.20
CBPE	O	O	$\mathbf{25.34}_{\pm 0.39}$	$\mathbf{16.79}_{\pm 0.18}$	$\mathbf{8.56}_{\pm 0.09}$	$\mathbf{30.51}_{\pm 0.69}$	$\mathbf{20.54}_{\pm 0.59}$	$\mathbf{9.97}_{\pm 0.12}$	**59.74**

CBPE models the right-hand side terms, $p(y \mid s, z, m)$ and $p(z)$, by optimizing the evidence ELBO in Equation 1 with observational data. By conditioning only on the causal parents of Y, CBPE blocks back-door paths that could introduce spurious correlations into the prediction.

Algorithm 1 details the CBPE framework, initialized by identifying direct causes (M) and confounder proxies (W) via DAGMA [25]. The model jointly optimizes a VAE and an auxiliary predictor; the latter generates a surrogate $\tilde{y}$ to infer the latent confounder z since the true target is unavailable. Finally, inference predicts $\hat{y}$ by fusing embeddings of the PPG signal, causal attributes, and the inferred latent confounder via $p_\theta(y \mid s, z, m)$.

5 Experiments

Our experiments aim to address the following aspects: (1) We compare CBPE with baselines on SABP to evaluate its performance and verify its superiority. (2) We investigate how well CBPE mitigates group-wise performance disparities in the presence of bias caused by group imbalance and introduce a new evaluation metric for this purpose. (3) We analyze the impact of the latent variable Z on the performance of CBPE. We compare our method against a set of signal-based baseline models commonly used for blood pressure (BP) estimation, particularly those relying on PPG signals, such as SpectroResNet [15], MLPBP [3], and Conv1D [42]. The baselines also include transformer-based architectures widely adopted in signal modeling, such as TimeMixer [14] and HGCTNet [18], as well as CEVAE [33], a causal effect prediction model that treats all attributes as observed confounders without explicitly modeling their causal relationships.

5.1 Evaluation metrics

We compare the performance of CBPE with baseline methods for BP estimation. The evaluation is conducted at both the sample level and the group level. Here, we define sample-level loss (S-loss) and group-level loss (G-loss) as follows:

$$\mathcal{L}^{\mathrm{S}} = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}_i|, \quad \mathcal{L}^{\mathrm{G}} = \frac{1}{|\mathcal{G}|}\sum_{j=1}^{|\mathcal{G}|} \mathcal{L}_j^{\mathrm{S}},$$

where $\mathcal{L}^{\mathrm{S}}$ denotes the S-loss, N represents the total sample count, and y_i and $\hat{y}_i$ denote the true and predicted BP of sample i, respectively. To compare the bias of estimators relative to the dataset distribution, we divided the data into four groups (hypotension, normal, prehypertension, and hypertension) denoted by $\mathcal{G}$ based on SBP and DBP ranges following GloGen [2]. G-loss ($\mathcal{L}^{\mathrm{G}}$) is computed as the mean of S-loss values across $\mathcal{G}$, where $\mathcal{L}_j^{\mathrm{S}}$ denotes S-loss for group j.

Group Loss Consistency metric. In the medical domain, maintaining consistency in performance across different groups is crucial. However, S-loss and G-loss alone are insufficient to evaluate performance imbalances among groups. For instance, if the Normal group exhibits very low loss while the Hypo group records high loss, conventional metrics fail to capture this imbalance. To address this issue, we propose the Group Loss Consistency (GLC) metric, which evaluates models based on both overall performance and the balance of loss distribution across groups. The GLC metric is defined as:

$$\mathrm{GLC} = \mathcal{L}^{G}\left(2 + \frac{\sum_{j=1}^{|\mathcal{G}|} p_j \log p_j}{\log |\mathcal{G}|}\right), \text{where } p_j = \mathcal{L}_j^{\mathrm{S}} / \sum_{k=1}^{|\mathcal{G}|} \mathcal{L}_k^{\mathrm{S}} \quad (4)$$

The normalized entropy term captures the imbalance in the loss distribution across groups. Lower GLC values indicate reduced overall loss and improved balance across groups. The entropy component penalizes inconsistent group losses, ensuring equitable outcomes in biomedical applications. GLC converges to the G-loss when group losses are perfectly balanced.

5.2 Quantitative analysis

We compute S-loss and G-loss for SBP and DBP and sum them to obtain BP S-loss and BP G-loss. As shown in Table 2, CBPE outperforms all baselines, with a notably larger improvement in BP G-loss, indicating superior robustness and reduced group bias. Its lowest GLC further confirms consistent and equitable BP estimation by modeling latent confounders that capture unobserved BP-related variations.

5.3 Ablation and Qualitative Analysis

Effectiveness of the latent variable Z in CBPE. To evaluate the role of the latent variable Z, we compare Conv1D-based variants using different combinations of attributes. As shown in Table 3, incorporating attributes enhances performance; however, relying solely on causally identified ones yields limited gains. CBPE achieves the

Table 3: Ablation study showing the impact of attribute information and latent causal features. NC, C, and LC denote Non-Causal, Causal, and Latent Causal attributes, respectively. Using only causal attributes (Conv1D-SC) leads to degraded performance, likely because the model relies on a reduced amount of attribute information. In contrast, CBPE yields more balanced predictions over BP groups than combining causal with non-causal attributes.

Method	PPG Signal	Attributes NC	C	LC	BP S-loss	BP G-loss	GLC
Conv1D	O	X	X	X	30.70	37.39	73.56
Conv1D-SA	O	O	O	X	27.39	34.56	66.75
Conv1D-SC	O	X	O	X	28.74	35.97	69.40
CBPE	O	X	O	O	**25.34**	**30.51**	**60.64**

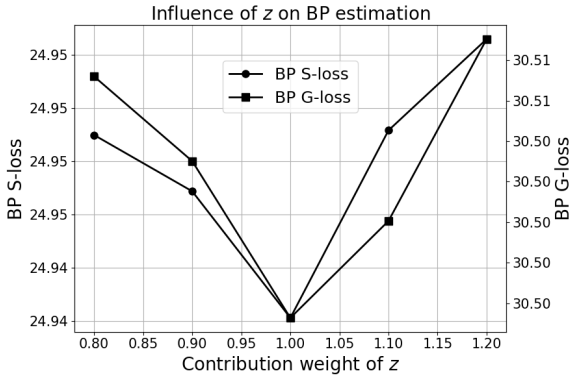

Figure 4: Influence of the latent confounder embedding on blood pressure estimation. Given PPG embedding, attributes, and latent z as input to the final classifier for predicting Y, we measure performance by varying the contribution of z via a scalar weight. Both BP S-loss and BP G-loss are minimized at the original weight, while either attenuating or amplifying z degrades performance, indicating that the latent causal feature z is beneficial but must contribute in a balanced way relative to the other attributes.

best results by jointly modeling causal attributes and latent confounders, demonstrating improved accuracy and robustness across patient groups.

To assess the contribution of the latent variable Z in CBPE, we compare Conv1D-based variants: Conv1D (PPG only), Conv1D-SA (PPG with all observed attributes), and Conv1D-SC (PPG with only causal BP-related attributes identified via causal discovery). Table 3 shows that adding attributes improves performance over Conv1D. However, Conv1D-SC yields lower accuracy than Conv1D-SA, indicating that the discovered causal set omits some BP-relevant information. CBPE, which infers hidden confounders to model latent causal features, performs best, implying that ignoring unobserved confounding or omitting relevant attributes lowers accuracy. When the causal structure is uncertain, using all features can help; however, explicitly combining identified causal attributes with latent confounders, as CBPE does, yields the most accurate BP estimates.

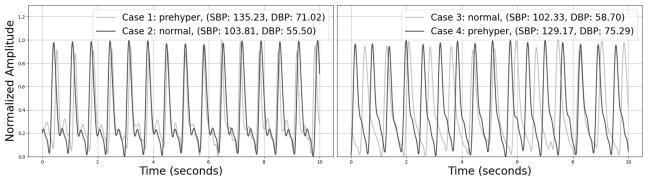

Figure 5: PPG signal comparison across cases. Case 1 and Case 2 exhibit similar PPG signals, with a difference of approximately 0.168. In contrast, Case 3 and Case 4 show a larger difference of about 0.339, indicating lower signal similarity. On average, the PPG distance difference is 0.317. The difference between PPG signals is measured as the mean absolute difference between two sequences.

Table 4: Comparison of BP loss for each case. Cases 1 and 2 have similar signals but different attributes, so the estimator should focus on attributes for accurate BP prediction. Conversely, Cases 3 and 4 share attributes but differ in signals, requiring the estimator to rely on signal variations. Corresponding PPG signals are shown in Figure 5.

		TimeMixer	CBPE
Sim. signal & Diff. Att.	Case 1	24.19	**2.63**
	Case 2	23.26	**1.23**
Diff. signal & Same Att.	Case 3	25.90	**19.75**
	Case 4	15.63	**12.37**

Figure 4 shows that increasing the weight on z raises sample-level loss and group-level loss, while reducing the weight also degrades performance, indicating that maintaining a contribution around the original scale leads to optimal performance. This suggests that the latent causal feature z indeed provides useful information for blood pressure estimation, yielding the most robust predictions when its influence is kept in balance with the other inputs rather than being overly suppressed or amplified.

Case analysis. Figure 5 and Table 4 highlight CBPE's ability to disentangle information from PPG signals and patient attributes. When signal information is ambiguous (Cases 1–2), CBPE leverages attribute cues for accurate prediction; when attributes are similar, but signals differ (Cases 3–4), it relies on signal variations. This adaptive weighting across modalities demonstrates CBPE's robustness and superior performance in complex scenarios.

6 Conclusion and Discussion

We introduce SABP, a multimodal benchmark for blood pressure (BP) prediction that integrates high-quality PPG waveforms with comprehensive demographic and clinical attributes. Building on this dataset, we propose CBPE, a causal framework that combines signal, tabular, and text modalities while inferring latent confounders to enhance both accuracy and group consistency. We further introduce the Group Loss Consistency (GLC) metric to jointly evaluate fairness and accuracy over BP groups. Extensive experiments show that CBPE achieves superior BP prediction performance and the lowest GLC, demonstrating equitable and reliable BP estimation across diverse patient groups.

Acknowledgments

This work was supported by Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean government [24ZD1140, Regional Industry ICT Convergence Technology Advancement and Support Project in Daegu-GyeongBuk (Medical) and 25ZD1140, Regional Industry ICT Convergence Technology Advancement and Support Project in Daegu-GyeongBuk (Medical)], also supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2024-00457216).

References

[1] Flávio D Fuchs and Paul K Whelton. 2020. High blood pressure and cardiovascular disease. *Hypertension*, 75, 2, 285–292.

[2] Taero Kim, Hyeonjeong Lee, Minseong Kim, Kwang-Yong Kim, Kyu Hyung Kim, and Kyungwoo Song. 2024. Glogen: ppg prompts for few-shot transfer learning in blood pressure estimation. *Computers in Biology and Medicine*, 183, 109216. doi:https://doi.org/10.1016/j.compbiomed.2024.109216.

[3] Bin Huang, Weihai Chen, Chun-Liang Lin, Chia-Feng Juang, and Jianhua Wang. 2022. Mlp-bp: a novel framework for cuffless blood pressure measurement with ppg and ecg signals based on mlp-mixer neural networks. *Biomedical Signal Processing and Control*, 73, 103404.

[4] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2, 11, 665–673.

[5] Wenqian Ye, Guangtao Zheng, Xu Cao, Yunsheng Ma, and Aidong Zhang. 2024. Spurious correlations in machine learning: a survey. *arXiv preprint arXiv:2402.12715*.

[6] Yongbo Liang, Zhencheng Chen, Guiyong Liu, and Mohamed Elgendi. 2018. A new, short-recorded photoplethysmogram dataset for blood pressure monitoring in china. *Scientific data*, 5, 1, 1–7.

[7] Charles Carlson, Vanessa-Rose Turpin, Ahmad Suliman, Carl Ade, Steve Warren, and David E Thompson. 2020. Bed-based ballistocardiography: dataset and ability to track cardiovascular parameters. *Sensors*, 21, 1, 156.

[8] Nicolas Aguirre, Edith Grall-Maës, Leandro J Cymberknop, and Ricardo L Armentano. 2021. Blood pressure morphology assessment from photoplethysmogram and demographic information using deep learning with attention mechanism. *Sensors*, 21, 6, 2167.

[9] Weinan Wang, Pedram Mohseni, Kevin L. Kilgore, and Laleh Najafizadeh. 2023. Pulsedb: a large, cleaned dataset based on mimic-iii and vitaldb for benchmarking cuff-less blood pressure estimation methods. *Frontiers in Digital Health*, 4. doi:10.3389/fdgth.2022.1090854.

[10] Alistair Johnson et al. 2016. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3, (May 2016), 160035. doi:10.1038/sdata.2016.35.

[11] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

[12] Gloria Martínez, Newton Howard, Derek Abbott, Kenneth Lim, Rabab Ward, and Mohamed Elgendi. 2018. Can photoplethysmography replace arterial blood pressure in the assessment of blood pressure? *Journal of clinical medicine*, 7, 10, 316.

[13] Tharindu Fernando, Sridha Sridharan, Simon Denman, and Clinton Fookes. 2022. Split 'n' merge net: a dynamic masking network for multi-task attention. *Pattern Recognition*, 126, 108551. doi:https://doi.org/10.1016/j.patcog.2022.108551.

[14] Shiyu Wang, Haixu Wu, Xiaoming Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Y Zhang, and Jun Zhou. 2024. Timemixer: decomposable multiscale mixing for time series forecasting. *arXiv preprint arXiv:2405.14616*.

[15] Gašper Slapničar, Nejc Mlakar, and Mitja Luštrek. 2019. Blood pressure estimation from photoplethysmogram using a spectro-temporal deep neural network. *Sensors*, 19, 15, 3420.

[16] Sergio González, Wan-Ting Hsieh, and Trista Pei-Chun Chen. 2023. A benchmark for machine-learning based non-invasive blood pressure estimation using photoplethysmogram. *Scientific Data*, 10, 1, 149.

[17] Ziyi Liu, Yiming Zhang, and Congcong Zhou. 2024. Bigru-attention for continuous blood pressure trends estimation through single channel ppg. *Computers in Biology and Medicine*, 168, 107795.

[18] Zeng-Ding Liu, Ye Li, Yuan-Ting Zhang, Jia Zeng, Zu-Xian Chen, Ji-Kui Liu, and Fen Miao. 2024. Hgctnet: handcrafted feature-guided cnn and transformer network for wearable cuffless blood pressure measurement. *IEEE Journal of Biomedical and Health Informatics*.

[19] Lei Liu, Yuan-Ting Zhang, Wenyan Wang, Yifan Chen, and Xiaorong Ding. 2023. Causal inference based cuffless blood pressure estimation: a pilot study. *Computers in Biology and Medicine*, 159, 106900. doi:https://doi.org/10.1016/j.compbiomed.2023.106900.

[20] Donald B Rubin. 2005. Causal inference using potential outcomes. *Journal of the American Statistical Association*, 100, 469, 322–331. eprint: https://doi.org/10.1198/016214504000001880. doi:10.1198/016214504000001880.

[21] Judea Pearl. 2015. Trygve haavelmo and the emergence of causal calculus. *Econometric Theory*, 31, 1, 152–179. doi:10.1017/S0266466614000231.

[22] Jonas Peters, Dominik Janzing, and Bernhard Schlkopf. 2017. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press. ISBN: 0262037319.

[23] Peter Spirtes and Richard Scheines. 1993. *Causation, Prediction, and Search*. Vol. 81. (Jan. 1993). ISBN: 978-1-4612-7650-0. doi:10.1007/978-1-4612-2748-9.

[24] Jiji Zhang. 2008. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172, 16, 1873–1896. doi:https://doi.org/10.1016/j.artint.2008.08.001.

[25] Kevin Bello, Bryon Aragam, and Pradeep Ravikumar. 2022. Dagma: learning dags via m-matrices and a log-determinant acyclicity characterization. *Advances in Neural Information Processing Systems*, 35, 8226–8239.

[26] Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. 2018. Dags with no tears: continuous optimization for structure learning. *Advances in neural information processing systems*, 31.

[27] Alain Hauser and Peter Bühlmann. 2012. Characterization and greedy learning of interventional markov equivalence classes of directed acyclic graphs. *The Journal of Machine Learning Research*, 13, 1, 2409–2464.

[28] Karren Yang, Abigail Katcoff, and Caroline Uhler. 2018. Characterizing and learning equivalence classes of causal dags under interventions. In *International Conference on Machine Learning*. PMLR, 5541–5550.

[29] Jinsung Yoon, James Jordon, and Mihaela Van Der Schaar. 2018. Ganite: estimation of individualized treatment effects using generative adversarial nets. In *International conference on learning representations*.

[30] Ahmed M Alaa and Mihaela Van Der Schaar. 2017. Bayesian inference of individualized treatment effects using multi-task gaussian processes. *Advances in neural information processing systems*, 30.

[31] Uri Shalit, Fredrik D Johansson, and David Sontag. 2017. Estimating individual treatment effect: generalization bounds and algorithms. In *International conference on machine learning*. PMLR, 3076–3085.

[32] Claudia Shi, David Blei, and Victor Veitch. 2019. Adapting neural networks for the estimation of treatment effects. *Advances in neural information processing systems*, 32.

[33] Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. 2017. Causal effect inference with deep latent-variable models. *Advances in neural information processing systems*, 30.

[34] Stefan Feuerriegel et al. 2024. Causal machine learning for predicting treatment outcomes. *Nature Medicine*, 30, 4, 958–968.

[35] Pulakesh Upadhyaya, Kai Zhang, Can Li, Xiaoqian Jiang, Yejin Kim, et al. 2023. Scalable causal structure learning: scoping review of traditional and deep learning algorithms and new opportunities in biomedicine. *JMIR Medical Informatics*, 11, 1, e38266.

[36] Chaochao Lu, Yuhuai Wu, José Miguel Hernández-Lobato, and Bernhard Schölkopf. 2021. Invariant causal representation learning for out-of-distribution generalization. In *International Conference on Learning Representations*.

[37] Jivat Neet Kaur, Emre Kiciman, and Amit Sharma. 2022. Modeling the data-generating process is necessary for out-of-distribution generalization. *arXiv preprint arXiv:2206.07837*.

[38] Chang Liu, Xinwei Sun, Jindong Wang, Haoyue Tang, Tao Li, Tao Qin, Wei Chen, and Tie-Yan Liu. 2021. Learning causal semantic representation for out-of-distribution prediction. In *Advances in Neural Information Processing Systems*. M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, (Eds.) Vol. 34. Curran Associates, Inc., 6155–6170. https://proceedings.neurips.cc/paper_files/paper/2021/file/310614fca8fb8e5491295336298c340f-Paper.pdf.

[39] Hyung-Chul Lee, Yoonsang Park, Soo Bin Yoon, Seong Mi Yang, Dongnyeok Park, and Chul-Woo Jung. 2022. Vitaldb, a high-fidelity multi-parameter vital signs database in surgical patients. *Scientific Data*, 9, 1, 279. ISBN: 2052-4463. doi:10.1038/s41597-022-01411-5.

[40] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. 2016. *Causal inference in statistics: A primer*. John Wiley & Sons.

[41] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. 2016. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78, 5, 947–1012.

[42] Oded Schlesinger, Nitai Vigderhouse, Danny Eytan, and Yair Moshe. 2020. 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)*. IEEE, 1135–1139.