Chapter 3

Never-Miss-A-Beat: A Physiological Digital Twins Framework for Cardiovascular Health

3.1 Digital Twins Relating PPG and ECG Sensing: Motivation and Problem Formulation

Under the umbrella of physiological digital twins as described in Chapter 1.1.3, the contribution of this chapter focuses on a particular application of a digital twin in healthcare for monitoring a person's cardiac activity. Cardiovascular disease (CVD) is the leading cause of mortality worldwide, accounting for 18.6 million deaths in 2019 [121] and clinical data suggest that the susceptibility to outcomes of COVID-19 is strongly associated with CVD [105]. Thus, the ability to consistently and accurately monitor cardiac activity is extremely important. Two commonly used cardiac sensing modalities that we are already familiar with are electrocardiogram (ECG) [49] and photoplethysmogram (PPG) [7]. ECG and PPG each have strengths and limitations in clinical practice: most notably, the clinical gold standard of ECG is monitored sporadically (commonly for 30-second intervals and, even with specialty devices, rarely over two weeks) and requires a

user's attention and cooperation as summarized in Table 2.1, while PPG can be monitored continuously but has a significantly smaller clinical knowledge base than ECG and tends to be noisy (although denoising is possible [175]). The ability to leverage the advantages of both technologies could have major impacts on the healthcare system, leading to easier everyday health monitoring.

ECG and PPG represent different but closely related physiological quantities, but how they are related is not well understood quantitatively. In this work, we have made some early-stage efforts toward understanding this relationship depending on age groups and cardiovascular conditions, as well as individualized nature. Thus, developing an explainable cardio-physiological digital twin model provides an excellent opportunity for monitoring a person's cardiac activities.

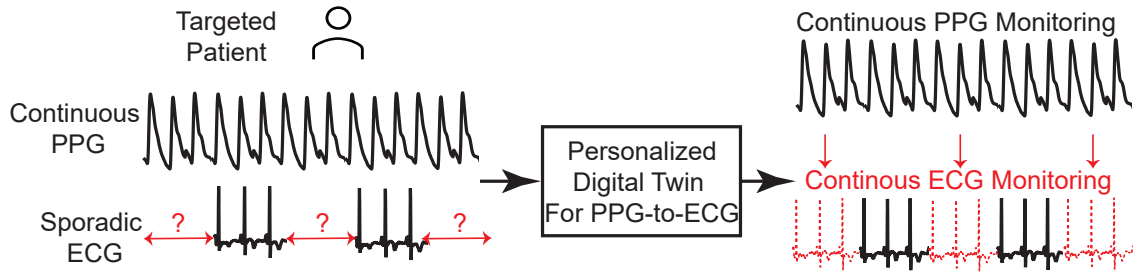


Figure 3.1: Our goal in this work is to build a personalized digital twin model for a targeted patient with his/her limited sporadic paired PPG and ECG cycles, such that his/her ECG can be faithfully and continuously inferred from the continuous PPG measured by the daily wearable devices.

More specifically, we pose the following question: is it possible to leverage continuous PPG monitoring, build a digital twin model to establish a patient's personalized PPG and ECG relationship during sporadic ECG sensing sessions, and use digital twins to infer continuous ECG waveforms? Through smart interpolation or extrapolation enabled by the digital twins, we can support continuous ECG monitoring, never missing

a beat, as illustrated in Fig. 3.1. This can be particularly valuable for helping patients and physicians capture details of cardiac events that are not commonly exhibited during a patient's clinical visits. By monitoring the digital twin that represents the real-time heart electrical activities and blood circulation in cyberspace, a centralized server or cardiologists can identify sudden cardiac risks so that high-risk populations can receive early medical intervention and even prevent premature mortality.

3.2 Related Background

	Dataset Split			
	TBME-RR [77]	MIMIC-III [74]	BIDMC [113]	SPCup 2015 [169]
Zhu et al. [175]	8-min from each of the 42 patients; 80%:20% train/test split from each patient	Three 5-min sessions from each of the 103 patients; 2:1 train/test split from each patient	n/a	n/a
Tian et al. [144]	Same as Zhu et al. [175]	Three 5-min sessions from each of the 33 patients; 80%:20% train/test split from each patient	n/a	5-min from each of the 13 participants; 80%:20% train/test split from each patient
Li et al. [92]	n/a	Same as Tian et al. [144]	8-min from each of the 53 patients; 80%:20% train/test split from each patient	Same as Tian et al. [144]
Vo et al. [154]	n/a	8-min from each of the 276 patients; 80%:20% train/test split from each patient	n/a	n/a
Chiu et al. [31]	n/a	n/a	Didn't mention	n/a

Table 3.1: A research review of the dataset and its split method used by the emerging technologies for ECG waveform inference from continuous PPG.

In recent years, researchers have begun to bridge the ECG-PPG knowledge gap by modeling the relationship between these two signals [92, 143, 144, 154, 174, 175], including the work presented in Chapter 2. Among those works, Vo et al. [154] used randomly selected 8-minute-long PPG-ECG signals from 276 patients in the MIMIC-II database [52] to analyze their models. Zhu et al., 2019 [174]; Zhu et al., 2021 [175]; Tian et al., 2020 [143]; Tian et al., 2021 [144], and Li et al. [92] evaluated their mod-

els on PPG-ECG signal pairs taken from the MIMIC-III database. In all of these works, results are provided in which 80% of the pairs were used for training and validation, while the remaining 20% of the data were used for model testing. Table 3.1 provides an overview of the emerging technologies for ECG waveform inference from continuous PPG. Because this split is carried out over all the data, these results are discussed in an average/generic sense that is out of context with precision healthcare. On the other hand, Zhu et al., 2021 [175] also trained subject-specific models in which the data from a particular subject is used to train a personalized model for analyzing the ECG reconstruction performance from PPG. Even though the subject-specific results are more relevant for precision healthcare, they also used the 80-20% training-testing splits to train their model. In real-world application scenarios, subject-specific data may be more scarce, which may cause these models to break down.

3.3 Methodology

3.3.1 Backbone Model for ECG Inference from PPG

Among the prior arts [92, 143, 144, 154, 174, 175] dedicated to the PPG-based ECG inference problem summarized in Chapter 3.2, the pilot study [174] first proved the feasibility of inferring ECG waveforms from PPG sensors by relating the two signals in the discrete cosine transform (DCT) domain using linear regression. Despite its computational efficiency, the DCT method [174, 175] lacks enough data representation power to faithfully reproduce ECG from PPG signals when the morphology of ECG waves becomes complex due to cardiovascular complications. Neural networks, with strong expressive

power and high structural flexibility, are also adopted to solve this problem [92, 154]. However, the computational cost of deep neural networks hinders their widespread deployment in practical applications. Also, black-box large neural network models are difficult for cardiologists to interpret and be receptive by the results. To strike a balance between the accuracy of ECG inference and computational resources in real-world scenarios, we first start with the dictionary-learning-based framework XDJDL proposed in Chapter 2 as a backbone model for PPG-to-ECG inference that provides a proper solution: compared to the DCT method, it improves the data representation with versatile and adaptive models; and it can perform efficiently in terms of power consumption and computational cost [6, 93]. The neural network based backbone models will be proposed and evaluated in Chapter 3.6 and Chapter 3.7.

Here is a summary and recapitulation of the key points in the XDJDL model that we adopt as the backbone model. Two dictionaries, $\mathbf{D}_p \in \mathbb{R}^{d\times k_p}$ and $\mathbf{D}_e \in \mathbb{R}^{d\times k_e}$, are learned jointly to estimate sparse representations ($\mathbf{A}_p \in \mathbb{R}^{k_p\times N}$ and $\mathbf{A}_e \in \mathbb{R}^{k_e\times N}$) for PPG and ECG datasets $\mathbf{X}_p \in \mathbb{R}^{d\times N}$ and $\mathbf{X}_e \in \mathbb{R}^{d\times N}$, respectively. Each column of $\mathbf{X}_p$ and $\mathbf{X}_e$ is denoted as $\mathbf{p}_i \in \mathbb{R}^{d\times 1}$ and $\mathbf{e}_i \in \mathbb{R}^{d\times 1}$, representing one PPG/ECG signal pair from the same cardiac cycle. Simultaneously, a linear transformation $\mathbf{W}$ is learned to map the sparse codes from the PPG to the ECG. The problem of solving for $\mathbf{D}_p$, $\mathbf{D}_e$, and $\mathbf{W}$ is formalized in Eq. (3.1).

$$\begin{aligned} \min_{\mathbf{D}_e,\mathbf{A}_e,\mathbf{D}_p,\mathbf{A}_p,\mathbf{W}} \quad & \|\mathbf{X}_e - \mathbf{D}_e\mathbf{A}_e\|_F^2 + \alpha\|\mathbf{X}_p - \mathbf{D}_p\mathbf{A}_p\|_F^2 + \beta\|\mathbf{A}_e - \mathbf{W}\mathbf{A}_p\|_F^2 \\ \text{s.t.} \quad & \|\mathbf{a}_{p,j}\|_0 \le t_p, \quad \|\mathbf{a}_{e,j}\|_0 \le t_e. \end{aligned} \tag{3.1}$$

The first two terms in Eq. (3.1), coupled with the constraints on the upper limits for sparsity, are used to learn the dictionary pair and sparse PPG and ECG representations iteratively by the two-step optimization strategy explained in Chapter 2.3.2 that is composed of sparse coding and dictionary update, while the third term in the equation facilitates learning the mapping between the two sparse domains simultaneously. In this way, the representation error in the first two terms and the mapping error in the third term are minimized. Fig. 3.2 summarizes the learning procedure of XDJDL.

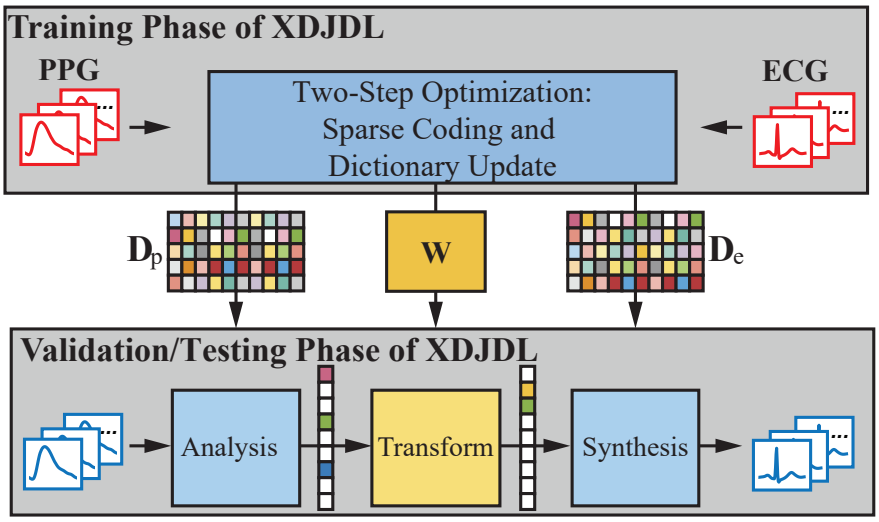


Figure 3.2: The XDJDL model proposed in Chapter 2 is adopted here as the backbone model for ECG inference from PPG. The dictionary pair $\mathbf{D}_p$, $\mathbf{D}_e$, and the linear mapping $\mathbf{W}$ are learned during the training phase, which are later applied to infer ECG from PPG in the validation or testing phase.

3.3.2 Transfer Learning for Building Precision Healthcare Digital Twins

Considering a group of people with abundant paired PPG and ECG signals from their in-hospital stay or annual physical examinations, we denote their corresponding PPG and ECG datasets as $\mathbf{X}_p \in \mathbb{R}^{d\times N}$ and $\mathbf{X}_e \in \mathbb{R}^{d\times N}$, respectively. Each column of $\mathbf{X}_p$

and $\mathbf{X}_e$ represents one PPG/ECG signal pair from the same cardiac cycle. Given $\mathbf{X}_p$ and $\mathbf{X}_e$, we can learn a **generic digital twin model** to simulate the PPG-to-ECG mapping. The XDJDL backbone model we adopt from Chapter 2 has shown that this *group-based* model (referred to as the generic digital twin model in this chapter) can be applied to predict the future ECG waveforms well from the PPG waveforms of people in the same group.

In this chapter, we consider a more practical scenario in which we would like to perform continuous ECG monitoring for a new target participant who only provides sporadic short (mostly 30-second segments) PPG/ECG paired signals acquired from his/her wearable devices like the Apple Watch [138], AliveCor [76], Zio patch [39], and Empetica E4 watch [43]. We denote the corresponding PPG and ECG datasets as $\mathbf{T}_p \in \mathbb{R}^{d\times M}$ and $\mathbf{T}_e \in \mathbb{R}^{d\times M}(M \ll N)$, respectively. Our aim in this work is to propose a method to fully utilize the sporadic data of the new participant, so that a **precision healthcare digital twin model** can be learned for the specific participant to infer and monitor his/her ECG from PPG wearable devices.

To address the challenge of data scarcity from the target participant, we propose to transfer the knowledge inherited in the generic digital twin model learned from the training participants with abundant PPG/ECG recordings from in-hospital stays or annual examinations, so that the generic digital twin can be refined and tailored to the new participant. In this study, we learn the healthcare digital twin model in the following training modes:

1. ***Transfer Learning Mode (Proposed)***: As visualized in Fig. 3.3(a), during the trans-

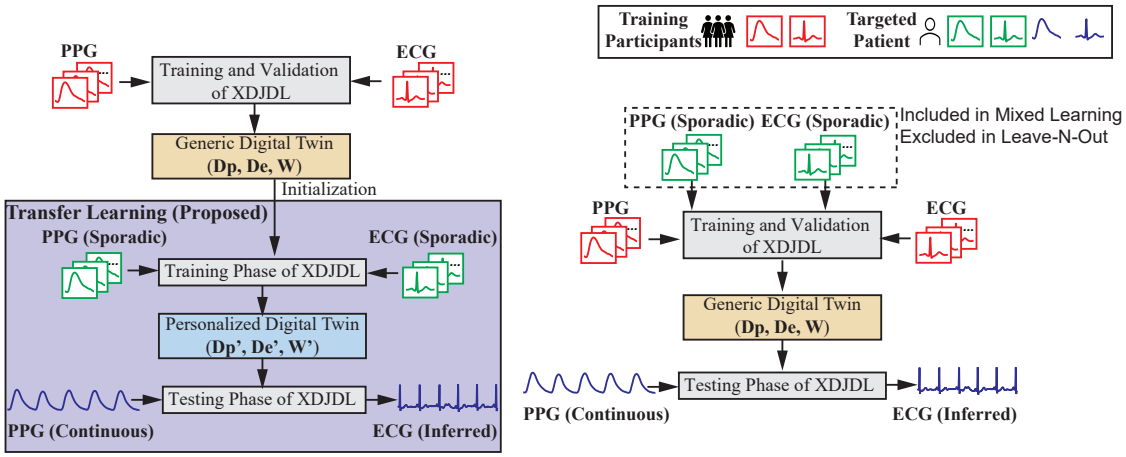


Figure 3.3: Flowcharts for (a) proposed transfer learning and (b) baseline comparisons including mixed learning and leave-N-out training scenarios. For the proposed transfer learning mode, a generic digital twin model is initially trained and validated using data from training participants, yielding the paired dictionaries $\mathbf{D}_p$, $\mathbf{D}_e$, and a linear transform $\mathbf{W}$. This model is then refined to be a personalized digital twin ($\mathbf{D}'_p$, $\mathbf{D}'_e$, and $\mathbf{W}'$) with the sporadic ECG and PPG pairs of the target patient. In the baseline comparisons, the generic digital twin is learned solely from the data of training participants in the leave-N-out mode and additional sporadic pairs from the target patient are used for training and validation in the mixed learning mode.

fer learning phase, the generic digital twin model learned from training participants serves as the initialization model. This is followed by continued training of the model on sporadic PPG/ECG pairs from the target participants, which updates the generic model variables $\mathbf{D}_p$, $\mathbf{D}_e$, and $\mathbf{W}$ to $\mathbf{D}'_p$, $\mathbf{D}'_e$, and $\mathbf{W}'$. These updates result in the proposed personalized digital twin model tailored to the target participants for precision healthcare.

2. ***Mixed Learning Mode (Baseline Comparison 1)***: As illustrated in Fig. 3.3(b), on top of using the long PPG/ECG paired recordings from the training participants, sporadic PPG/ECG pairs from the target patient are also included to learn the generic digital twin model $\mathbf{D}_p$, $\mathbf{D}_e$, and $\mathbf{W}$.

Compared to the transfer learning mode, the mixed learning mode requires model training from scratch with mixed data from training and target participants, which can be time-consuming and not realistic if the training data is not accessible. While in transfer learning mode, the generic digital twin model is used in a plug-and-play form that does not require the data from the training participants to retrain the model.

3. ***Leave-N-Out Mode (Baseline Comparison 2)****:* As displayed in Fig. 3.3(b), in this mode, we apply the generic digital twin model learned solely from the training participants to the new target patient. This mode provides the baseline performance without making use of the sporadic data from the target patients and reveals the adaptation capability of the generic digital twin model to unseen participants.

3.3.3 Testing Modes for ECG Inference

To detect symptoms of underlying heart conditions (like elevated heartbeat [4]) early for proper intervention, continuous long-term ECG monitoring is critical to pick up on subtle deviations from a person's normal ECG patterns. Discontinuous ECG signals may not fully capture critical deviating behavior which can lead to a wrong evaluation, deteriorating the effectiveness of treatments [19]. For this reason, once the digital twin model is learned, we present two testing modes in our analysis for addressing the issue of discontinuous ECG monitoring in realistic situations: interpolation and extrapolation.

1. ***Interpolation Mode (Illustrated in Fig. 3.4(a))****:* Suppose we have two short pairs of PPG/ECG signals with some time interval in between from the target participant

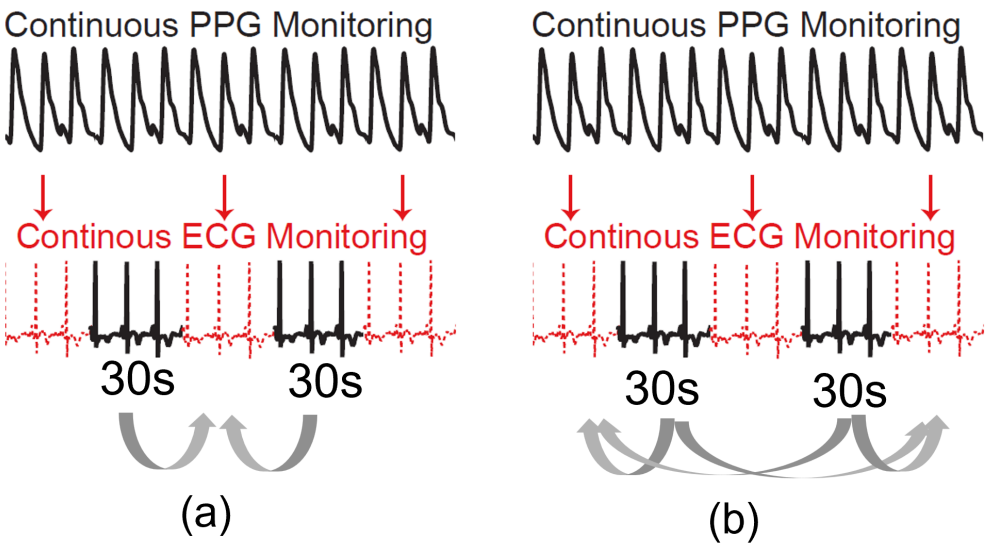


Figure 3.4: The two testing modes that we examine for the learned digital twin model. (a) Interpolation mode where we can "rewind" the ECG during its detachment between two sporadic time stamps that contain known paired PPG and ECG signals. (b) Extrapolation mode where we can check the past ECG or predict the future ECG.

and we aim to 'interpolate' the ECG waveforms from the continuous PPG signal acquired between the two sporadic time stamps. This interpolation mode corresponds to realistic situations where the participant wishes to detach the ECG nodes from his/her body for some time. The ECG information before detachment and after the reattachment can be used to "rewind time" to reconstruct the signal that was lost during the detached period.

2. ***Extrapolation Mode (Illustrated in Fig. 3.4(b)):*** Suppose we have two sporadic short pairs of PPG/ECG signals from the target participant and we aim to 'extrapolate' the ECG waveforms from the continuously acquired PPG signal before and after the two sporadic time stamps. This extrapolation mode corresponds to realistic situations where a medical practitioner wants to know what the ECG signal looked like in the past or to predict what will happen in the future. In the former

case, physiological abnormalities that otherwise would have been missed may be detected. In the latter case, preventative measures can be taken should the predicted future signal display physiological abnormalities that could lead to health concerns.

3.4 Experimental Results Using XDJDL as The Backbone For The Personalized Digital Twin Model

3.4.1 Dataset

Medical Information Mart for Intensive Care III (MIMIC-III) [74] is a large database comprised of health information related to patients admitted to the intensive care unit at the Beth Israel Deaconess Medical Center in Boston, Massachusetts. Timestamped bedside vital sign measurement is provided for each of the $53,423$ patient hospital admissions.

The analysis in this study is performed on a subset of data from the MIMIC-III database that was collected using the methodology outlined as follows. Patients that had paired lead II ECG and PPG signals in the record were selected from the waveform database and were linked to their patient profiles (sex, disease, etc) according to the subject IDs. Of these signals, only those of high quality and belonged to patients with specific cardiovascular/non-cardiovascular diseases were retained for analysis. Cardiovascular diseases were chosen from the list of "diseases of the circulatory system" based on the ICD-9 codes of the patients and the following cardiovascular diseases are included in the collected dataset: atrial fibrillation, myocardial infarction, cardiac arrest, congestive heart failure, hypotension, hypertension, and coronary artery disease. For non-cardiovascular

diseases, we selected sepsis, pneumonia, gastrointestinal bleed, diabetic ketoacidosis, and altered mental status under other categories of ICD-9 codes. The result was a set of 127 subjects as displayed in Fig. 3.5 with the age distributions of the cardiovascular disease and non-cardiovascular disease subjects. Each subject has three 5-minute sessions of paired ECG and PPG recordings which were collected within a few hours of each other. To differentiate this dataset from the mini-MIMIC-33 dataset evaluated in Chapter 2, we denote it as the *mini-MIMIC-127* dataset.

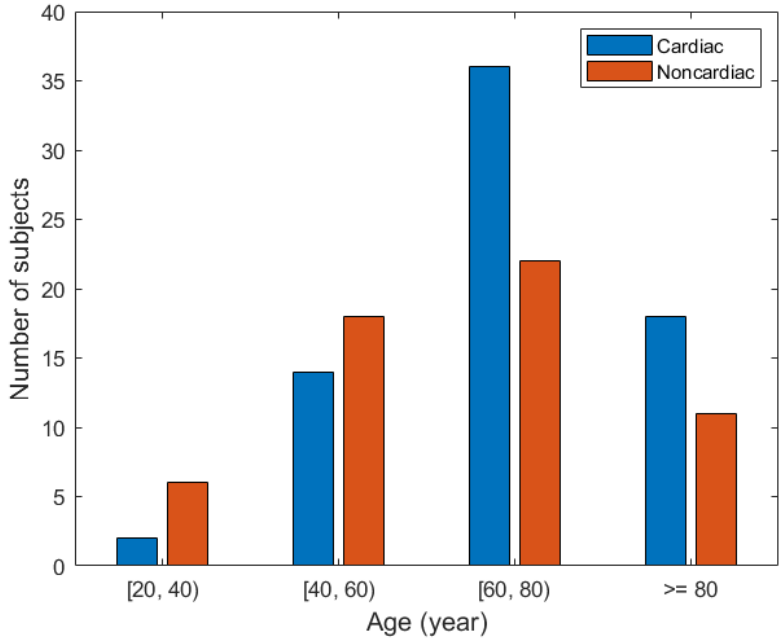


Figure 3.5: Distribution of the 127 patients collected from the MIMIC-III database in different age groups and disease types (mini-MIMIC-127 dataset). Within each age group, the patients with cardiovascular-related diseases are marked in blue on the left, and the patients with non-cardiovascular-related diseases are marked in orange on the right.

3.4.2 Hyperparameters Selection

In the XDJDL framework described in Eq. (3.1), dictionary sizes for PPG and ECG signals are important hyperparameters to be chosen for good data representation. In this

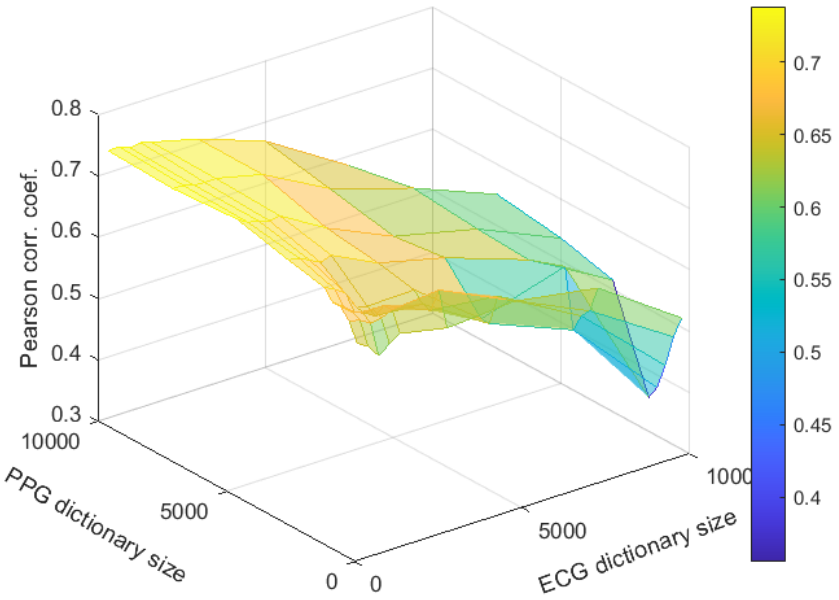


Figure 3.6: The validation performance in terms of Pearson correlation coefficient with respect to different combinations of PPG and ECG dictionary sizes.

section, we explain how we select the best sizes of the PPG and ECG dictionaries by examining their impact on the performance of the ECG reconstruction in terms of the Pearson coefficient. Different combinations of dictionary sizes are used to train the XD-JDL models with training data from the first two sessions of each training participant. The trained models are later evaluated on the validation set built from the third session of each training participant to select the proper size of the PPG and ECG dictionary pair.

From Fig. 3.6, we observe that given the same ECG dictionary size, the Pearson coefficient in the validation set improves and becomes saturated as the PPG dictionary size grows towards 10000. The trend of convergence suggests potential model overfitting. Another observation is that given the same PPG dictionary size, the performance remains almost unchanged and deteriorates as the ECG dictionary size increases. Hence, using fewer atoms for the ECG dictionary is a good choice. The experimental results indicating

that the number of atoms in a PPG dictionary needs to be much greater than the number of atoms in an ECG dictionary suggest that there are more detailed differences among PPG signals than ECG signals in the collected dataset.

This phenomenon that PPG needs far more atoms than ECG can be counterintuitive at first glance. Because from frequency analysis, people may find that ECG has more high frequency components and thus needs more atoms to be represented. But if we view ECG as the source and PPG as the downstream signal, according to information theory, we know that the entropy of the system will increase as the information flows from the heart to the peripheral vasculature after the processing of all the blood vessels along the way. As a result, PPG contains more subtlety and nuances and needs more atoms to present. One example that reinforces this assumption is that during severe hemorrhage (blood volume loss) caused by trauma injury, ECG contains fewer useful features for early detection of hemorrhage until irreversible harm or cardiovascular collapse but PPG senses this extreme medical situation sooner and is often used as an important bio-marker to detect blood loss in its early stage [30, 117, 118].

3.4.3 Performance of ECG Inference

We split the overall dataset with 127 patients into four groups according to their health-related physical attributes (age and disease type). These groups include 16 cardiac young patients (age less than 60 with cardiac diseases), 54 cardiac old patients (age greater than or equal to 60 with cardiac diseases), 24 noncardiac young patients (age less than 60 with noncardiac diseases), and 33 noncardiac old patients (age greater than or equal to 60

with noncardiac diseases). In this way, the generic digital twin model corresponding to each attribute group can be learned separately and applied to the target patients with the same attribute.

For each group, three patients are randomly selected as the target participants and the data from the rest of the patients in this group are used for training and validation. The training set is composed of the first two sessions from each patient and the validation set consists of the last session from the same patient. Thus, the current data split is training:validation:testing = 6:3:1 on average. This corresponds to the realistic setting of building a precision healthcare digital twin model with the generic digital twin learned from a large portion of patients to be applied to a few target patients.

For the *interpolation test mode*, the first 45 cycles (approximating a 30-second segment) from the first session and the last 45 cycles from the last session of the target participants are regarded as the known sporadic pairs for either transfer learning or mixed learning. The second session of the target participants is used to evaluate the interpolation performance. For the *extrapolation test mode*, the first and last 45 cycles from the second session of the target participants are regarded as the known sporadic pairs for either transfer learning or mixed learning. The first and last sessions are used to evaluate the extrapolation performance. It is worth noting that each participant only has three 5-min sessions in a sequence collected at most within a few hours, meaning the interpolation and extrapolation results in this work are for a relatively short period. For longer time window results, such as daily or weekly, a preliminary performance evaluation is shown in Chapter 3.5.2.

We use the Pearson correlation coefficient (ρ) and the relative root mean squared

Error (rRMSE) to evaluate the morphological fidelity of the inferred ECG $\hat{e}$:

$$\rho = \frac{(\boldsymbol{e} - \mu[\boldsymbol{e}])^{\mathrm{T}}(\hat{\boldsymbol{e}} - \mu[\hat{\boldsymbol{e}}])}{\|\boldsymbol{e} - \mu[\boldsymbol{e}]\|_2 \ \|\hat{\boldsymbol{e}} - \mu[\hat{\boldsymbol{e}}]\|_2}, \tag{3.2}$$

$$\text{rRMSE} = \frac{\|\boldsymbol{e} - \hat{\boldsymbol{e}}\|_2}{\|\boldsymbol{e}\|_2}, \tag{3.3}$$

where $\boldsymbol{e}$ denotes the reference ECG cycle, and $\mu[\cdot]$ represents the element-wise average of a vector.

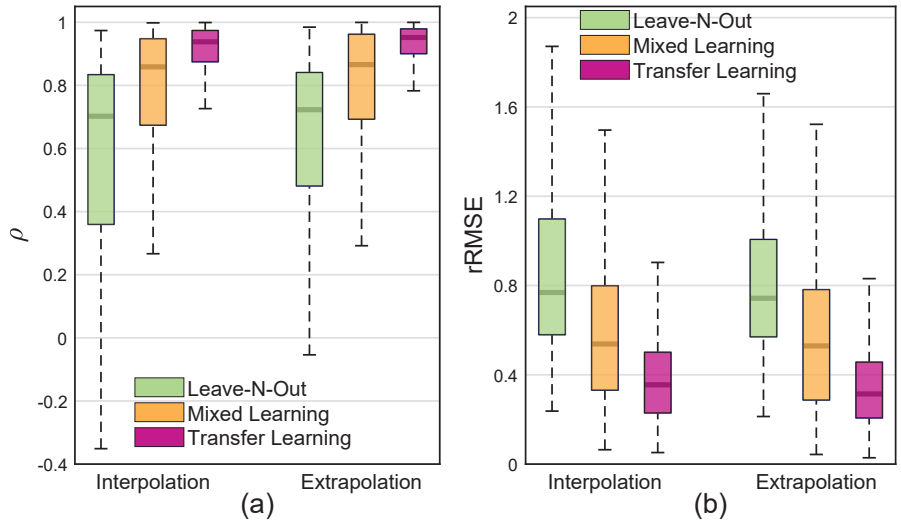


Figure 3.7: Statistical distribution of (a) Pearson correlation coefficient (ρ) and (b) rRMSE for the inferred ECG signals in both interpolation and extrapolation testing modes using different training modes (leave-N-out, mixed learning, and transfer learning).

Fig. 3.7 depicts the overall distribution comparison of the reconstruction performance summarized in the boxplots using XDJDL as the PPG-to-ECG inference model. Each boxplot is composed of the results from all groups. We observe that the medians and spreads of ρ and rRMSE improve from leave-N-out mode to mixed learning mode, and transfer learning mode achieves the best result in both interpolation and extrapolation

testing scenarios. Specifically, the medians of ρ in the interpolation testing mode are 0.70, 0.86, and 0.94 across the three training modes, respectively, while those in the extrapolation testing mode are 0.72, 0.87, and 0.95, respectively. The median rRMSE values in the interpolation testing mode are 0.77, 0.54, and 0.36 across the three training modes, respectively, while those in the extrapolation testing mode are 0.74, 0.53, and 0.31, respectively. Analysis of these boxplots suggests that the transfer learning mode can both interpolate and extrapolate ECG signals for the target participants from their sporadic PPG/ECG pairs with high fidelity, indicating the effectiveness of our proposed method in learning the precision healthcare digital twin.

	Interpolation		Extrapolation	
	ρ	rRMSE	ρ	rRMSE
Cardiac young group				
Transfer learning	**0.90 (0.09)**	**0.42 (0.21)**	**0.91 (0.11)**	**0.40 (0.24)**
Mixed learning	0.76 (0.26)	0.62 (0.33)	0.82 (0.22)	0.54 (0.29)
Leave-N-out	0.67 (0.23)	0.76 (0.26)	0.73 (0.17)	0.71 (0.21)
Cardiac old group				
Transfer learning	**0.95 (0.07)**	**0.28 (0.17)**	**0.95 (0.06)**	**0.28 (0.16)**
Mixed learning	0.66 (0.39)	0.69 (0.41)	0.60 (0.45)	0.76 (0.45)
Leave-N-out	0.58 (0.40)	0.82 (0.34)	0.61 (0.37)	0.81 (0.32)
Noncardiac young group				
Transfer learning	**0.87 (0.11)**	**0.49 (0.23)**	**0.88 (0.13)**	**0.48 (0.28)**
Mixed learning	0.78 (0.25)	0.57 (0.30)	0.74 (0.30)	0.62 (0.36)
Leave-N-out	0.32 (0.56)	1.07 (0.54)	0.34 (0.55)	1.06 (0.52)
Noncardiac old group				
Transfer learning	**0.92 (0.14)**	**0.37 (0.42)**	**0.95 (0.05)**	**0.28 (0.14)**
Mixed learning	0.80 (0.24)	0.52 (0.35)	0.89 (0.17)	0.39 (0.26)
Leave-N-out	0.59 (0.28)	0.85 (0.35)	0.66 (0.26)	0.76 (0.30)
Overall				
Transfer learning	**0.90 (0.11)**	**0.40 (0.28)**	**0.92 (0.10)**	**0.36 (0.23)**
Mixed learning	0.75 (0.30)	0.60 (0.35)	0.76 (0.33)	0.59 (0.38)
Leave-N-out	0.52 (0.43)	0.90 (0.42)	0.56 (0.42)	0.86 (0.40)

Table 3.2: Experimental results from each group and overall result from all groups for the inferred ECG in terms of the mean and the standard deviation (in parenthesis) of Pearson coefficient (ρ) and rRMSE.

In addition to the overall statistical distribution, Table 3.2 lists the ECG inference performance in terms of the mean and standard deviation of Pearson coefficient (ρ) and rRMSE for each group along with the overall results for all groups. The results in each group are consistent with the overall results as shown in Fig. 3.7 and the last three rows of Table 3.2: leave-N-out sets the baseline performance, mixed learning improves it with the target participant's sporadic data mixed in the training phase, and the proposed transfer learning mode further boasts an improved ECG inference performance. The only exception is in the cardiac old group where the mixed learning and leave-N-out achieve comparable performance in the extrapolation testing mode. This could be due to that the cardiac old group is the largest group (54 people in total), and the weight of the target participant is relatively small in the mixed-learning, making it comparable to the leave-one-out case. Another observation is that, except for the noncardiac young group, in the remaining three groups, the leave-N-out training mode can achieve reasonably fair reconstruction performance with a Pearson coefficient ρ of at least 0.58 and as high as 0.73. Since leave-N-out is the most challenging case, with the target patient's data totally unseen in the training phase, its acceptable quality of reconstruction indicates that separating patients into groups of similar attributes is helpful to achieve good reconstruction performance for people belonging to the same group given the current dataset. This generalization capability in an even larger dataset needs further validation, and more attributes can be considered, such as ethnic, different hospitals, etc.

Fig. 3.8 shows three visualization examples comparing reconstructed ECG signals to their reference ECG signals. In Fig. 3.8(a), the leave-N-out mode infers the ECG of this patient from the knowledge learned from all the training patients in the same group

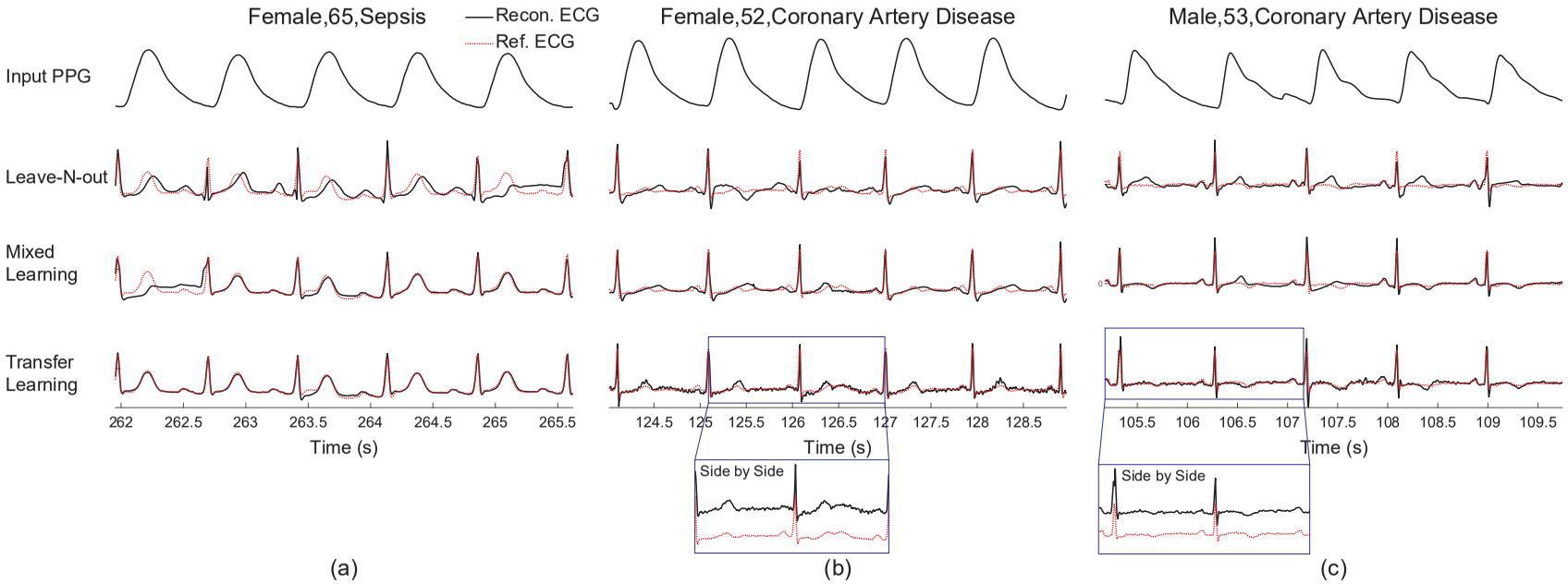


Figure 3.8: Qualitative comparison of the ECG signals inferred in different modes. Examples are from (a) a 65-year-old female with sepsis, (b) a 52-year-old female with coronary artery disease, and (c) a 53-year-old male with coronary artery disease. From top to bottom: the input PPG signal from which the ECG is inferred, results from the leave-N-out mode, results from the mixed learning mode, and results from the transfer learning mode. At the bottom of (b) and (c), we also provide a side-by-side view of specific cycles for illustration.

(noncardiac old). The first four cycles are reconstructed with high similarity to the reference ECG, with the T-wave slightly shifted in time. Mixed learning incorporates the target patient's sporadic ECG and PPG pairs in the training set, thus the last four cycles show improved reconstruction performance compared to the leave-N-out case, though a glitch still appears in the inference of the first cycle. Transfer learning further improves the reconstruction performance with all inferred cycles matching the reference signal. In Fig. 3.8(b), we show a side-by-side view in the blue box comparing the inferred ECG using the transfer learning method to the reference ECG signal. The third cycle (second cycle in the blue box) is slightly different from the typical ECG waveform of this patient with an extra ascending and descending slope before the T-wave. From the side-by-side view in the blue box, transfer learning can recover this variant well. In Fig. 3.8(c), the ECG waveform of a participant with coronary artery disease is displayed. This patient's ECG waveform typically contains an obviously inverted T-wave, though the inversion is milder in the first cycle of the highlighted blue box. Nevertheless, the transfer learning model is able to accurately capture both the more and less pronounced inversion characteristics. From the illustrations in Fig. 3.8(b) and (c), the ECG variation of the target patients is captured well in the transfer learning mode but not in the mixed learning mode, suggesting that it is more useful to inherit the knowledge from a generic digital twin and then fine-tune it with the target patient's data.

3.5 Discussions for XDJDL-based Personalized Digital Twin Model

3.5.1 Results Based on PPG Segmentation Scheme

In Chapter 3.4, we have evaluated different training and testing modes for digital twins models based on the assumption that the timestamps of R peaks in the reference ECG signals are available to segment the paired ECG and PPG signals into cycles. In realistic settings, we may not have the reference ECG signal for segmentation. Thus, we consider a practical scenario of reconstructing the ECG from the "estimated cycles" of PPG that are segmented by the PPG onsets instead of the R peaks of ECG signals. We denote:

- R2R scheme: segmentation scheme based on R peaks of ECG as is used in Chapter 3.4;
- O2O scheme: segmentation scheme based on PPG onsets.

Due to the discrepancy between the detected locations of PPG onset and R peak of ECG from the same cycle, the "estimated" PPG/ECG cycles using the O2O scheme vary from those segmented by the R2R scheme. Thus, compared to the R2R scheme, in the O2O scheme, further ECG inference error results from 1) the time misalignment between the R peak of the inferred ECG and that of the reference ECG and 2) the reconstructed waveform error. To single out the error caused by 2), on top of the O2O scheme, we compensate for the time offset caused by 1) by shifting each inferred ECG cycle in time so that the reference and reconstructed ECG signals are matched according to their R peaks.

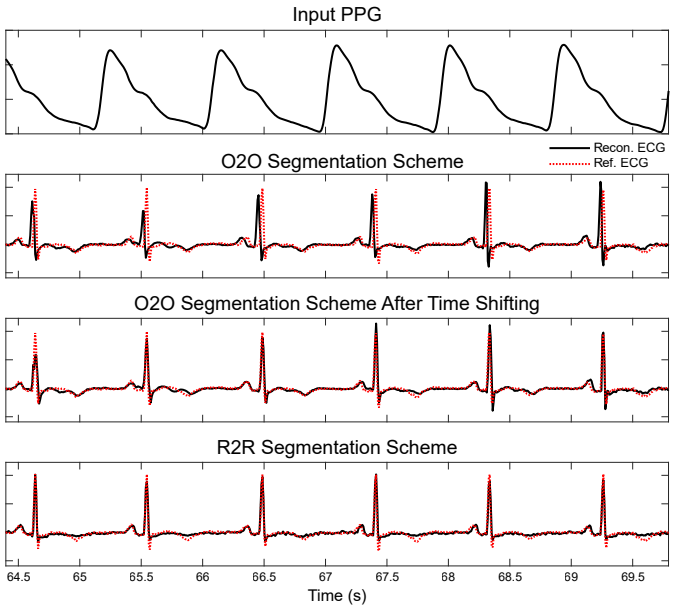


Figure 3.9: Qualitative comparison for different segmentation schemes. From top to bottom: the input PPG signal from which ECG is inferred, results from the O2O segmentation scheme, results after shifting the O2O inferred cycle in time to align the R peaks of the inferred ECG and the reference ECG, and results from the R2R segmentation scheme.

We denote the results after aligning R peaks of inferred ECG from the O2O scheme with the reference ECG as O2O'. One qualitative comparison example is shown in Fig. 3.9.

	ρ	rRMSE
Transfer learning (O2O)	0.45 (0.38)	1.08 (0.55)
Transfer learning (O2O')	0.75 (0.22)	0.75 (0.45)
Transfer learning (R2R)	**0.91 (0.10)**	**0.38 (0.25)**

Table 3.3: Comparison of different segmentation schemes for ECG inference presented in mean and standard deviation (in parenthesis) of Pearson coefficient (ρ) and rRMSE.

The overall comparison result is listed in Table 3.3. Compared to the R2R scheme, when using the O2O scheme, the average Pearson coefficient drops from 0.91 to 0.45, and the average rRMSE rises from 0.38 to 1.08. By compensating for the error from the misalignment of R peaks to only account for the waveform inference discrepancy, compared to O2O, O2O' improves the mean Pearson coefficient and the mean rRMSE to

0.75 and 0.75, respectively.

3.5.2 Performance Evaluation for Long Time Scale Data

This section aims to examine the performance of the personalized digital twins for ECG inference when the data are collected in a longer time window, e.g., during a week, in addition to the mini-MIMIC-127 dataset (Chapter 3.4.1) where each participant only has three 5-min sessions collected within a few hours. We self-collected the ECG and PPG data using consumer-grade sensors to test the temporal consistency of the personalized digital twins.

Self-collected Dataset:

One 27-year-old female subject participated in this week-long data collection (approved by University of Maryland IRB #1786518). This participant has not been diagnosed with any CVDs according to the most updated medical records. As shown in Table 3.4, 24 sessions for the subject at different times (morning, afternoon, and evening) of a day during a week were recorded. In each session, the participant was asked to hold the FDA-cleared EMAY portable ECG monitor (Model: EMG-10) to record the lead-I ECG. We measure the lead I ECG signal from the two hands, which is the easiest and most accessible way to use EMAY. Simultaneously, the index finger is placed in the CMS-50E pulse oximeter for PPG monitoring. The setup is shown in Fig. 3.10. It is worth noting that EMAY can only record a 30-second long ECG at a time, thus we asked the participant to hold it for 6 consecutive periods of ECG snapshots (3 minutes) in each session for longer recordings. To reduce the movement-induced artifacts and false diagnosis during

Subject 1							Year: 2022
Session		Session		Session		Session	
1	04-04, 11:53	8	04-06, 17:25	15	04-08, 21:37	22	04-11, 10:36
2	04-04, 16:08	9	04-06, 22:27	16	04-09, 10:31	23	04-11, 15:23
3	04-04, 20:38	10	04-07, 09:27	17	04-09, 15:39	24	04-11, 23:15
4	04-05, 09:04	11	04-07, 17:37	18	04-09, 23:31		
5	04-05, 15:15	12	04-07, 21:30	19	04-10, 09:01		
6	04-05, 21:38	13	04-08, 09:54	20	04-10, 15:12		
7	04-06, 08:58	14	04-08, 18:03	21	04-10, 21:19		

Table 3.4: The data collection time stamps for the participant during a week.

the recording, the participant was asked to sit comfortably and keep both hands on the desk as still as possible. The sampling rates of the EMAY ECG monitor and the PPG sensor are 250 Hz and 60 Hz, respectively. The PPG signal is upsampled to 250 Hz with spline interpolation. Then we preprocessed the signals using the same method as explained in Chapter 2.3.1.

Figure 3.10: Experimental setup for the self-collected PPG and ECG database. The CMS-50E pulse oximeter was measuring the PPG signal from the index finger and the EMAY was recording the lead-I ECG signal by connecting both hands to its metal electrodes.

Learning and Evaluation Schemes:

Given the attributes of the self-collected data, which includes young participants

with no known CVDs, we first learn a generic digital twin using the data from the 40 young patients from both cardiac and noncardiac groups in the mini-MIMIC-127 dataset from Chapter 3.4.1. With the generic digital twin model, we use the proposed transfer learning methodology (Chapter 3.3) to update it to a personalized model with the sporadic short paired PPG and ECG segments from the target participant.

We learn and evaluate the personalized digital twin in the following schemes:

- *Scheme (a) Interpolation & Extrapolation Within One Day*: Can we use paired PPG and ECG data from the morning and evening of each day to obtain the personalized digital twin and infer the afternoon data (i.e., interpolation within a day) and vice versa, use afternoon data to fine-tune the digital twin and infer the morning and evening data (i.e., extrapolation within a day)?

- *Scheme (b) Interpolation & Extrapolation Within Half A Week*: Can we use data from Day 3 morning & Day 6 evening to learn the personalized model to infer both interpolation case (sessions between Day 3 to Day 6) and extrapolation case (sessions from Day 1,2,7,8)?

- *Scheme (c) Interpolation Within A Week:* Can we use data from Day 1 morning & Day 8 evening to update the generic digital twin model to infer all the sessions in between?

ECG Inference Performance:

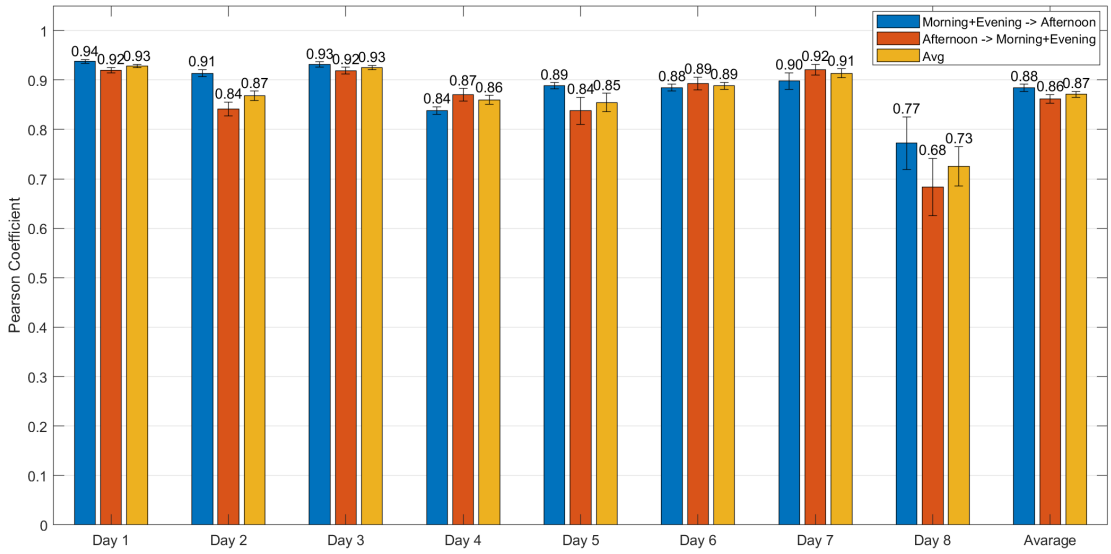
Table 3.5 summarizes the overall performance from each of the learning and evaluation schemes, excluding the training and validation sessions from Day 1 morning, Day 3 morning, Day 6 evening, and Day 8 evening for a fair comparison. We observe that

the personalized digital twin updated by the Day 1 morning and Day 8 evening data (Scheme (c)) achieves slightly better inference performance than the other two schemes, suggesting that Day 1 morning data is representative of the whole week's ECG-PPG relation for this target participant during the week of data collection.

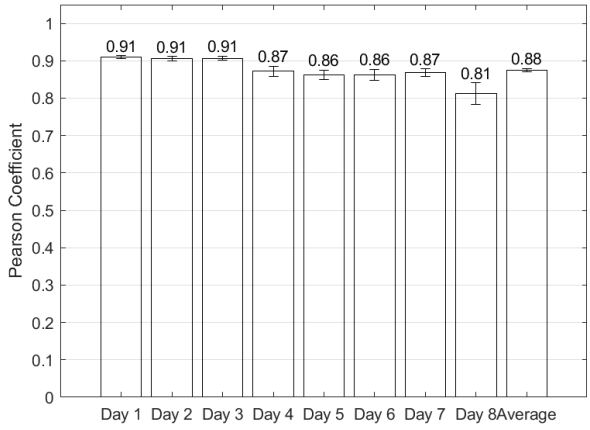
	Scheme (a)	Scheme (b)	Scheme (c)
ρ	0.87 (0.20)	0.88 (0.16)	0.88 (0.22)
rRMSE	0.49 (0.32)	0.49 (0.23)	0.44 (0.26)

Table 3.5: The personalized digital twin performance of different learning and evaluation schemes. Scheme (a) learns and evaluates the personalized digital twin daily, while Scheme (b) and Scheme (c) are conducted for data from several days to a week. Results are presented in means and standard deviations (in parentheses) of Pearson correlation coefficient ρ and rRMSE.

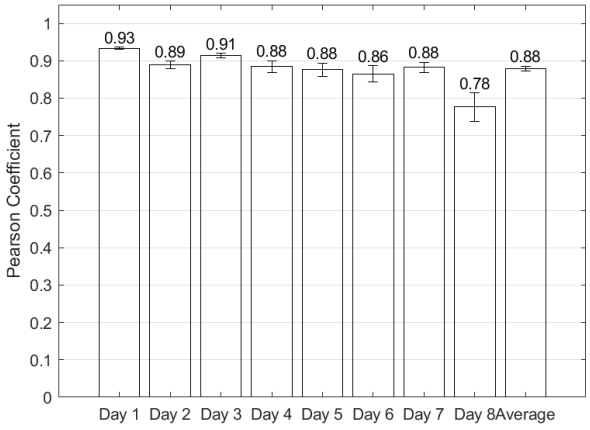
The breakdown of everyday performance in terms of the Pearson coefficient from the three schemes is shown in Fig. 3.11. The height of each bin shows the average correlation coefficient ρ of ECG reconstruction results from the overlapped test sessions of the three schemes each day. Each error bar corresponds to the 95% confidence interval that is calculated as $\pm 1.96\hat{\sigma}/\sqrt{N}$, where $\hat{\sigma}$ is the sample standard deviation and N is the sample size/number of ECG cycles. In Fig. 3.11a, the blue bar shows the results of the experiment for "interpolation within a day" that uses the morning and evening data to fine-tune a personalized digital twin to infer the afternoon ECG from the same day, and the red bar shows that for "extrapolation within a day" using the afternoon data to update the personalized digital twin to predict the morning and evening data, and the yellow bar is the averaged performance of "interpolation" and "extrapolation" modes. Comparing the results across the three schemes, we observe that the results are similar to each other from Day 1 to Day 7, and in more than a half of the days, Scheme (a) achieves slightly



(a)



(b)



(c)

Figure 3.11: The breakdown of everyday performance in terms of Pearson coefficient from the three schemes. (a), (b), and (c) show the results from Schemes (a), (b), and (c), respectively. Each bar represents the average Pearson Coefficient and each error bar represents the 95% confidence interval.

better performance than the other two schemes, suggesting that the inference within a day is more accurate than the prediction from several days apart. One exception/outlier is Day 8 when the averaged ρ of Scheme (a) is much lower than the other two schemes, especially the "extrapolation" mode of Scheme (a). This indicates that the afternoon data from Day 8 is not as representative to update the personal digital twin for the morning data of Day 8, compared to the morning data of Day 3 (i.e., Scheme (b)).

In a retrospect, the generalization performance of the personalized digital twin may be limited by a) the attribute difference between the training data and the self-collected data (ICU patients vs. healthy subjects) and b) the different leads of ECG signals collected in the training data and the self-collected data (lead II vs. lead I). Note that lead II is the most common and generally the best view because the placement of the positive electrode in Lead II views the wavefront of the impulse from the inferior aspect of the heart as it travels from the right arm (RA) towards the left leg (LL). Lead I ECG "views" the heart activity from the left arm (LA) to the right arm (RA) [79]. According to Einthoven's law, lead I + lead III = lead II, i.e., the sum of the potentials in lead I and lead III equals the potential in lead II. That may help explain that in the self-collected dataset, the amplitude of the R peak of ECG is generally less than 0.5mV while that of the training data is generally around 1mV to 2mV.

3.6 Using Neural Networks as The Backbone for ECG Inference from PPG to Build Digital Twins

In this section, we aim to improve the personalized digital twins with neural network based methods, which are more flexible for various transfer learning techniques than the XDJDL model as the backbone. A conditional variational autoencoder (CVAE) model is adopted here as the backbone model for PPG-to-ECG inference. Its capability of learning latent variables is suitable for manifesting the interpretability of the underlying physiological process relating PPG and ECG signals. Furthermore, in Chapter 3.7, a causal representation learning structure is proposed based on the CVAE architecture here for better explainability. To differentiate from the causal CVAE model that will be proposed in Chapter 3.7, we denote the CVAE model used in this section as the "vanilla CVAE" model.

3.6.1 A Retrospect: The Physiological Process Behind PPG and ECG Generation

In our previous work on PPG-to-ECG inference (Chapter 2), we have considered the ECG as the source signal and PPG as the downstream filtered signal and viewed it as an inverse engineering problem, as is shown in the yellow box of Fig. 3.12. However, if we take the full signal generation path into consideration as illustrated in the pink box of Fig. 3.12, we know that the myocardial activities (such as the impulse from the SA node) initiate the electrical signal in the heart. On the one hand, the varying electrical potentials

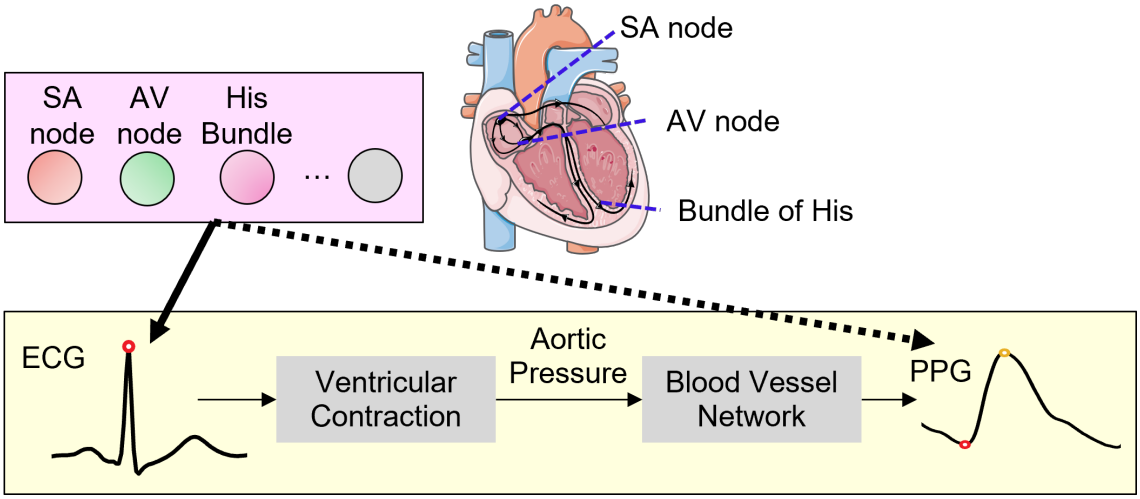


Figure 3.12: The ECG and PPG signal generation paths during heartbeats considering the originating impulses from the heart.

are captured by the skin electrodes of ECG sensors. On the other hand, the electrical pulse spreads in the heart, leading to the mechanical movements of the heart and the corresponding aortic pressure wave that later passes into the blood vessel network. The peripheral pulse wave is measured from the extremities with a PPG sensor, which received the light modulated by the transmissive and reflective interactions of the human skin. With this full physiological process in mind, we aim to consider the common factor, the heart activity, that generates both PPG and ECG into the picture and leverage the CVAE model to learn a latent variable $\mathbf{z}$ to represent this common source. The assumption we make with the vanilla CVAE is that this common source is Gaussian i.i.d. for all people. This is a relatively general assumption and we will see how to refine it in Chapter 3.7 with causal interpretation.

3.6.2 Conditional Variational Autoencoder (CVAE) for PPG-to-ECG Inference

To start with, we draw the connection with the previously proposed PPG-to-ECG methods, such as DCT-based [175], XDJDL-based [144] (Chapter 2), and autoencoder-based frameworks [92], before diving into the CVAE model. They all can be viewed to be designed to maximize the log of likelihood $P(Y|X,\Theta)$. This is because if we suppose $Y = \Theta(X) + z$, where $z \sim \mathcal{N}(0,\sigma^2)$, then $P(Y|X,\Theta) \sim \mathcal{N}(\Theta(X),\sigma^2)$ and the maximum log-likelihood problem can be translated to minimizing $\|\Theta(X) - Y\|^2$. In the DCT-based framework, X is the DCT feature from PPG and Y is that from ECG; in the XDJDL-based framework, X is the sparse representation for PPG and Y is that for ECG; and in the autoencoder-based framework, X is PPG and Y is ECG themselves.

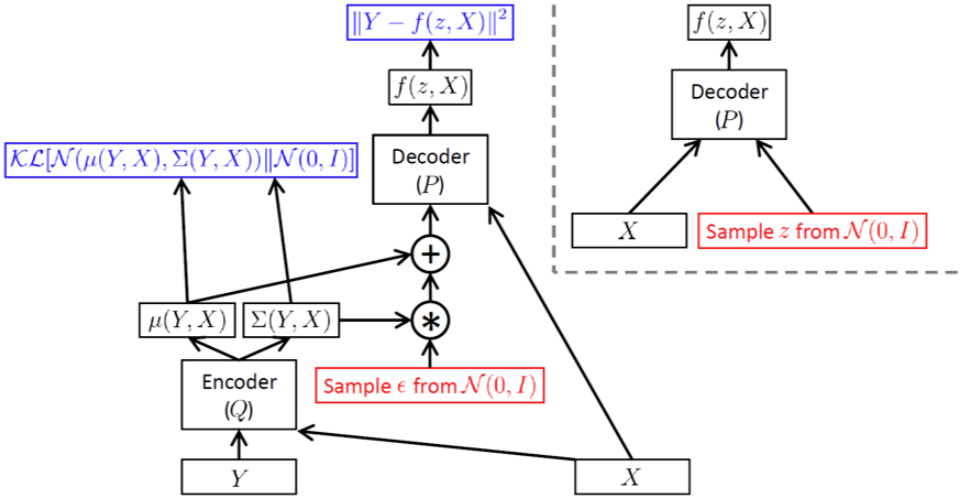


Figure 3.13: The left panel shows the CVAE structure implemented as a feed-forward neural network during the training process. The upper right panel shows the model at test time when we want to sample from $P(Y|X)$. The illustration is adopted from [41].

The CVAE structure illustrated in Figure 3.13 represents the core CVAE mathematical model in Equation (3.4). Instead of maximizing the log-likelihood on the left-hand

side of Equation (3.4), CVAE tries to optimize the surrogate objective function, the variational lower bound (ELBO) of the log-likelihood, which is the right-hand side of Equation (3.4). The first part of the ELBO can be regarded as the reconstruction accuracy, which is shown in the top blue box of Fig. 3.13. The second part of the ELBO is the KL divergence between the conditional distribution $Q(\mathbf{z}|Y,X)$ and $P(\mathbf{z}|X)$ represented in the leftmost blue box in Fig. 3.13, where $P(\mathbf{z}|X)$ is $\mathcal{N}(0,I)$ because the CVAE model assumes the latent variable $\mathbf{z}$ is sampled independently of X at the test time.

$$\begin{aligned} & logP(Y|X) - KL[Q(\mathbf{z}|Y,X)||P(\mathbf{z}|Y,X)] \\ & = E_{\mathbf{z}\sim Q(\cdot|Y,X)}[logP(Y|\mathbf{z},X)] - KL[Q(\mathbf{z}|Y,X)||P(\mathbf{z}|X)] \end{aligned} \tag{3.4}$$

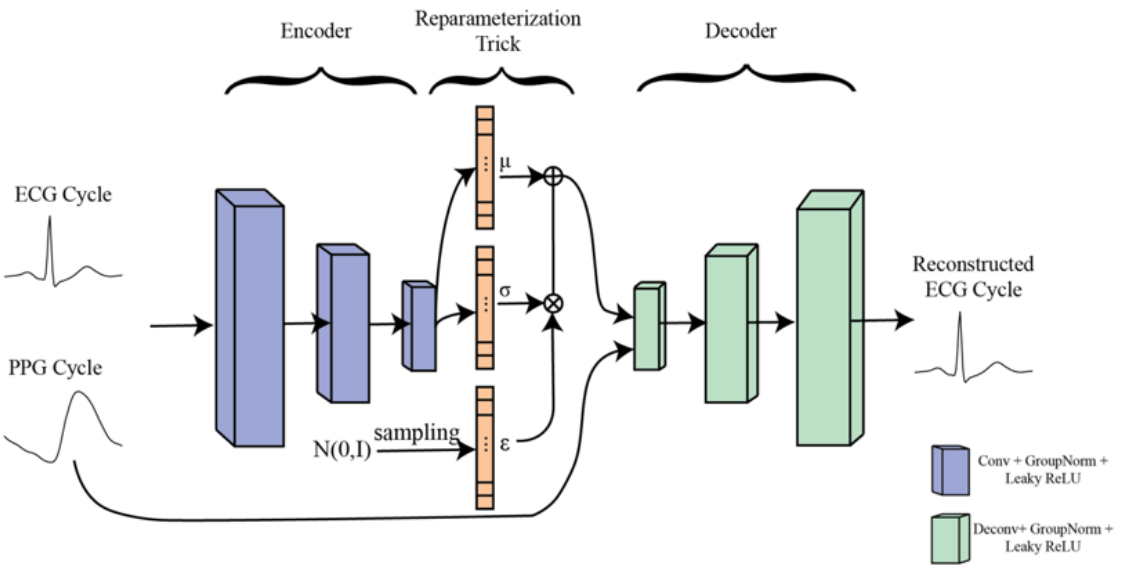


Figure 3.14: The vanilla CVAE model as the backbone for ECG inference from PPG.

Following the structure of CVAE, we use a convolutional neural network (CNN) to build it up. The overview of the model architecture is shown in Figure 3.14. We treat the ECG signal as the Y to be predicted and the PPG signal as the X on which the prediction is conditioned. The encoder and decoder are composed of a three-layer CNN,

respectively. Each layer starts with a convolution/deconvolution kernel (channel # 60, 40, and 40 and kernel size # 30, 15, and 5 in each convolution layer, and the parameter for the deconvolution layer is reversed), then a group normalization, followed by a LeakyReLU activation layer. The outputs of the encoder are the mean and variance for the latent variable, which is sampled using the reparameterization trick. Later on, the latent variable $\mathbf{z}$ will be concatenated with the PPG cycle as the label information to generate an ECG cycle. The size of the latent variable is chosen based on the best validation performance among 16, 32, and 64.

3.6.3 Transfer Learning to Build Personalized Digital Twin for Cardiovascular Monitoring

We repeat what has been proposed and done in Chapter 3.3 and Chapter 3.4 to build the personalized digital twin with vanilla CVAE rather than XDJDL as the backbone this time. As mentioned above, neural networks provide more options for transfer learning as it is flexible to fine-tune specific layers or add a few layers. We adopt three different fine-tuning methods: (a) tuning the first deconvolution layer in the decoder, (b) tuning all deconvolution layers in the decoder, and (c) tuning all parameters in the CVAE model.

Fig. 3.15 presents the comparison of the overall results between using XDJDL and CVAE as the backbone models for leave-N-out, mixed learning, and transfer learning with the three fine-tuning methods. The height of each bin shows the average correlation coefficient ρ or the rRMSE of ECG reconstruction results from both interpolation and extrapolation test modes of all participants. Each error bar corresponds to the 95%

	Interpolation		Extrapolation	
	ρ	rRMSE	ρ	rRMSE
Cardiac young group				
Leave-N-Out	0.70 (0.19)	0.73 (0.23)	0.75(0.14)	0.66 (0.16)
Mixed Learning	0.86 (0.12)	0.48 (0.16)	0.89 (0.12)	0.41 (0.20)
Transfer Learning (one layer)	0.92 (0.08)	0.38 (0.12)	0.95 (0.07)	0.29 (0.10)
Transfer Learning (encoder)	0.92 (0.07)	0.39 (0.13)	0.95 (0.07)	0.30 (0.11)
Transfer Learning (all parameters)	**0.94 (0.06)**	**0.32 (0.10)**	**0.96 (0.03)**	**0.27 (0.10)**
Cardiac old group				
Leave-N-Out	0.61 (0.26)	0.77 (0.19)	0.65 (0.26)	0.74 (0.20)
Mixed Learning	0.80 (0.24)	0.51 (0.29)	0.81 (0.27)	0.50 (0.29)
Transfer Learning (one layer)	0.91 (0.16)	0.33 (0.22)	0.93 (0.14)	0.32 (0.18)
Transfer Learning (encoder)	0.91 (0.16)	0.33 (0.21)	0.91 (0.18)	0.34 (0.20)
Transfer Learning (all parameters)	**0.95 (0.08)**	**0.26 (0.16)**	**0.95 (0.11)**	**0.27 (0.15)**
Noncardiac young group				
Leave-N-Out	0.43 (0.52)	0.91 (0.47)	0.47 (0.50)	0.88 (0.45)
Mixed Learning	0.89 (0.11)	0.42 (0.18)	0.85 (0.16)	0.48 (0.23)
Transfer Learning (one layer)	0.92 (0.05)	0.39 (0.11)	0.92 (0.08)	0.39 (0.13)
Transfer Learning (encoder)	0.92 (0.05)	0.40 (0.11)	0.91 (0.08)	0.41 (0.13)
Transfer Learning (all parameters)	**0.94 (0.05)**	**0.32 (0.11)**	**0.93 (0.08)**	**0.34 (0.15)**
Noncardiac old group				
Leave-N-Out	0.74 (0.15)	0.66 (0.19)	0.76 (0.13)	0.64 (0.17)
Mixed Learning	0.89 (0.14)	0.40 (0.23)	0.93 (0.10)	0.31 (0.17)
Transfer Learning (one layer)	0.94 (0.11)	0.31 (0.16)	0.96 (0.07)	0.27 (0.12)
Transfer Learning (encoder)	0.93 (0.12)	0.33 (0.17)	0.95 (0.07)	0.28 (0.12)
Transfer Learning (all parameters)	**0.95 (0.10)**	**0.26 (0.16)**	**0.97 (0.04)**	**0.23 (0.10)**

Table 3.6: The results using vanilla CVAE as the backbone model for the inferred ECG of each group in terms of the mean and the standard deviation (in parenthesis) of Pearson coefficient (ρ) and rRMSE.

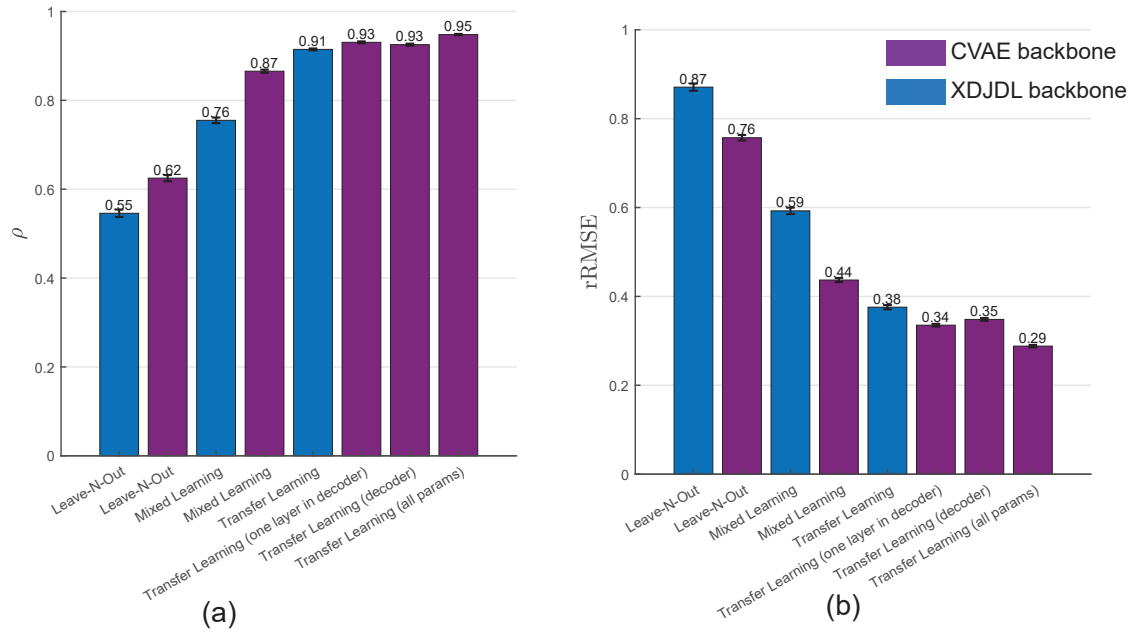


Figure 3.15: The overall performance comparison using XDJDL and CVAE as the backbone models for PPG-to-ECG inference in different training modes, including leave-N-out, mixed learning, and transfer learning. The error bars correspond to the 95% confidence intervals.

confidence interval that is calculated as $\pm 1.96\hat{\sigma}/\sqrt{N}$, where $\hat{\sigma}$ is the sample standard deviation and N is the sample size/number of ECG cycles. The breakdown of performance for each group of target participants is listed in Table 3.6. First, compared to Table 3.2 with XDJDL as the backbone model, we observe that there is an improvement in terms of the ECG reconstruction performance using vanilla CVAE as the backbone model in all training modes across all participant groups (Table 3.6) and overall participants (Fig. 3.15). Second, transfer learning with tuning all parameters achieves better performance than only tuning part of the parameters. In addition, tuning only the first layer in the decoder is almost comparable to tuning all the parameters. For practical applications, we may consider only tuning just one layer as this strikes a balance between the algorithm performance and computing resources.

3.7 Incorporating Causality into CVAE Model Based on Structural Causal Model (SCM)

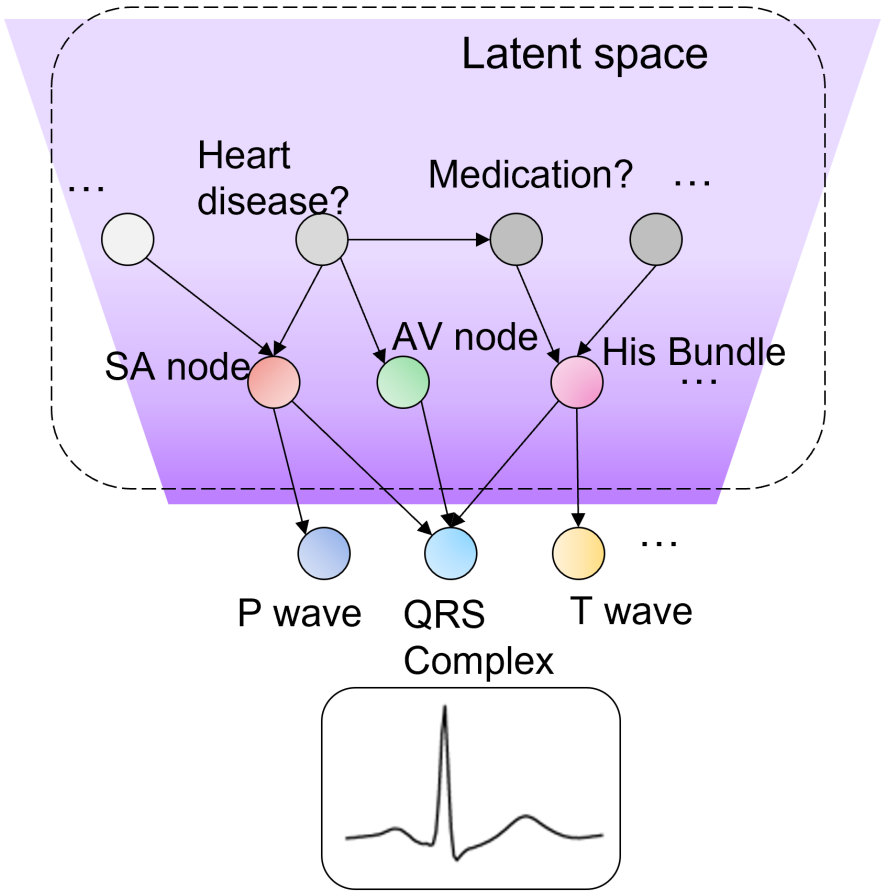


Figure 3.16: Illustration of causal representation learning for ECG inference. The factors that form a causal mechanism in the latent space are assumed to generate the higher dimensional data of the ECG waveform in the observed space. The arrow indicates that the parent node causes the child one.

In the previous vanilla CVAE model for PPG-to-ECG inference, we assume that the latent vector $\mathbf{z}$ representing the factors during the heart muscle mechanism to generate ECG signals conditioned on PPG is multivariate independent Gaussian for all people. In the realistic world, this assumption may be too general. To better fit our aim of building personalized digital twin model, in this section, we take one step further and propose

to learn a causal representation for generating ECG signals to improve the previous assumption. The key underlying assumption we make here is that the high dimensional observational data, which is the ECG signal in our case, is a manifestation of a lower dimensional set of factors AND those factors contain causal relationships among each other, such as the sample causal graph shown in Fig. 3.16, considering the physiological process of a heartbeat. Those factors that affect the ECG signal of each people may be personalized and more clinically interpretable rather than being a general i.i.d. multivariate Gaussian for all people. We aim to discover the proper causal representations in the sense that with the "do" operation to the latent representation factor, the higher dimensional observational data (e.g., ECG waveform) will be causally changed accordingly. The semantic/medical meaning of the nodes in the latent vectors may be subject-specific, and we will analyze them on a case-by-case basis.

3.7.1 Importance of Incorporating Causality into Machine Learning Algorithms and Structural Causal Model

With the fast development of big data and enhanced computational power, machine learning models, including deep learning models, have been growing fast in the past decade. In the healthcare field, they have been widely applied and have shown great predictive power, such as for disease classification [45, 62] and physiological signal sensing [15, 32]. However, good prediction performance indicating that there exists a statistical association between the input data and output labels does not necessarily imply causation between them [158]. For example, in [21], the authors aimed to predict the

probability of death for patients with pneumonia so that high-risk patients could be admitted to the hospital while low-risk patients were treated as outpatients. From their feature analysis, they found some counterintuitive relations between the input feature and the output prediction, e.g., if a patient has a history of asthma then the patient has a lower risk of death from pneumonia. This observation does not comply with the cause-and-effect in common sense and the reason could be that if a patient had asthma before, it is likely that the patient was treated once and thus has a lower possibility of death. Adding causal analysis to machine learning models would be tremendously useful to avoid the counterintuitive results and make the models more interpretable and it is drawing increasing attention nowadays to take the advantage of both fields [122].

Causal Directed Acyclic Graph (DAG) and Structural Causal Model (SCM):

Consider a set of random variables $X_1, ..., X_n$ building a DAG structure which is a graph with directed links between nodes but without directed cycles (acyclic). Note that Bayesian Network is a DAG where the joint probability distribution of the nodes (random variables) in the graph is $p(X_1, ..., X_n) = \prod_{i=1}^{n} p(X_i|\mathbf{PA}_i)$, i.e., a node is independent of its non-descendants given its parents. However, the general DAG that carries the Markov property of the conditional independence assumption is not enough to depict the quantitative causal relation among the nodes in the DAG that accounts for the generation of the data. SCM describes the causal mechanisms of a system with structural equations. A functional causal model is proposed in [112] to illustrate how the children vertices in the DAG are influenced by their parents in an ordering from the hypothesized cause-effect relations, i.e., $X_i := f_i(\mathbf{PA}_i, U_i), i = 1, ..., n$, where U_i represents arbitrary disturbance due to omitted factors that are mutually independent, f_i is a linear or nonlinear function,

and $\mathbf{PA}_i$ is the set of Markovian parents of X_i.

The linear structural equations model (SEM) [112] is a specialization of the functional causal model with generalized functional relation f_i, i.e., $X_i := \sum_{j \in \mathbf{PA}_i} A_{ji} X_j + U_i$. Suppose the linear adjacency matrix of SEM associated with the DAG structure is $\mathbf{A} \in \mathbb{R}^{n \times n}$, where A_{ij} represents the casual strength from Node i to Node j ($A_{ij} = 0$ if there is no causal edge from Node i to Node j). Then the linear SEM can be expressed in a matrix form as follows:

$$\mathbf{x} = \mathbf{A}^{\mathrm{T}} \mathbf{x} + \epsilon \tag{3.5}$$

where $\epsilon_i \sim \mathcal{N}(0, 1), i = 1, ..., n$ and $\mathbf{x} \in \mathbb{R}^{n \times 1}$. Useful properties of $\mathbf{A}$ are that: (1) $\mathbf{A}$ can be permuted into a strictly upper triangular matrix if the nodes in the DAG are strictly in causal ordering; (2) The ith column of $\mathbf{A}$ are the parents of the ith factor and the ith row of $\mathbf{A}$ are the children of the ith factor.

Notion of Do Intervention:

In [112], Pearl introduced the notion of "do(x)" for setting $X = x$ to distinguish it from the notion of pure "x" for observing $X = x$. In particular, the operation of do(x_j) means: (1) deleting the edges directed to the variable node X_j from $\mathbf{PA}_j$ in the DAG and the corresponding structural equation $x_j = f_i(\mathbf{PA}_j, u_j)$ in the SCM and (2) setting $X_j = x_j$ in the right-hand sides of the other equations of a causal structure in SCM. By investigating the mapping from x to $P(y|do(x))$ for all x, the causal effect of X on Y can be examined.

3.7.2 *Causal CVAE* Model for PPG-to-ECG Inference

On top of the vanilla CVAE model that assumes the learned latent factors in the latent vector are i.i.d. Gaussian (ϵ), we develop the causal CVAE model in this section as shown in Fig. 3.17. Instead of directly inputting the ϵ together with the PPG cycle into the decoder, we add a causal representation learning module after ϵ to learn the causal representation vector $\mathbf{z}$ based on the linear SEM via the *linear layer* and then pass $\mathbf{z}$ into the *DAG layer* validate that the causal mechanism holds for $\mathbf{z}$. This causal representation learning module is inspired by the work of CausalVAE proposed in [164]. The differences between CausalVAE [164] and our work in this section mainly lie in:

1. Our model is based on the conditional VAE while their method is proposed for VAE. This is not a trivial update especially when we are dealing with a different application scenario for ECG waveform inference from PPG;

2. The true causal label in [164] is assumed to be known and incorporated in training as a supervised problem, while that is unknown and complicated in our medical setting and is proposed to be estimated from the intervention experiment.

Here is the detailed design of the two additional layers in the casual representation learning module:

1. Linear layer: $\mathbf{z} = \mathbf{A}^{\mathrm{T}}\mathbf{z} + \epsilon = (\mathbf{I} - \mathbf{A}^{\mathrm{T}})^{-1}\epsilon$. This layer is designed based on the SEM in Eq. 3.5. The adjacent matrix $\mathbf{A}$ is learned during the training time to achieve the optimal causal representation $\mathbf{z}$;

2. DAG layer: $\mathbf{z} = f(\mathbf{A} \circ \mathbf{z}) + \epsilon$, where $\circ$ represents the element-wise multiplication

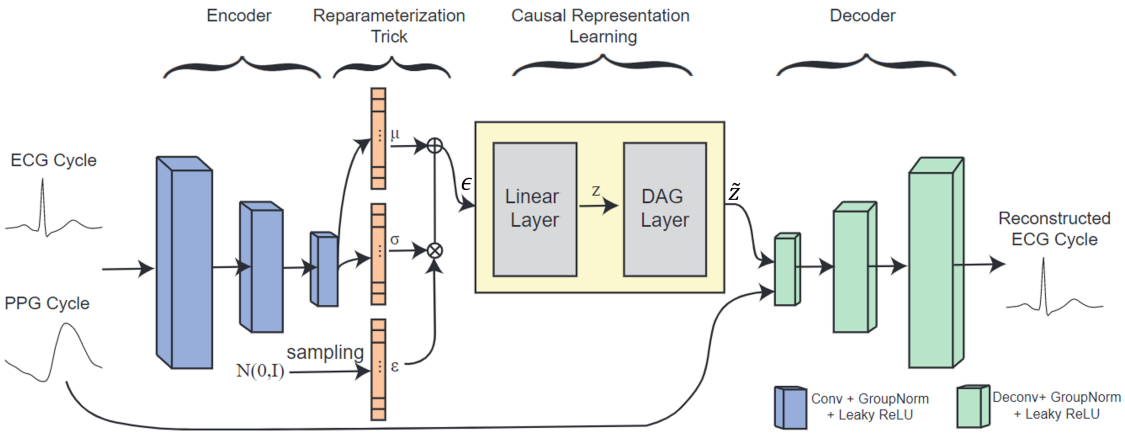


Figure 3.17: The proposed causal CVAE architecture. Compared to the vanilla CVAE structure in Fig. 3.14, the causal CVAE model incorporates the causal representation learning module that helps to learn the causal latent vector $\mathbf{z} = [z_1, ...z_n]^{\mathrm{T}} \in \mathbb{R}^n$ where z_i represents the ith node in the learned DAG.

of each column of $\mathbf{A}$ and $\mathbf{z}$. This layer resembles the SCM which depicts how children nodes are generated/influenced by their corresponding parental variables. f adds nonlinearity during training time. Note that this layer is necessary to conduct the intervention experiment that will be discussed later in Chapter 3.7.4.

Based on the architecture, the following loss functions are taken into account during the training process:

1. Acyclic enforcement on the DAG related adjacent matrix $\mathbf{A}$ [166]:

   $\mathcal{L}_{dagness} = \mathrm{tr}((\mathbf{I} + \frac{1}{n}\mathbf{A} \circ \mathbf{A})^n) - n$;

2. Enforcing $\mathbf{A}$ to be a non-zero matrix:

   $\mathcal{L}_{nonzero} = 1/\tanh(\frac{1}{n^2}\sum_{i=1}^{i=n}\sum_{j=1}^{j=n}|A_{i,j}| + \delta)$, where δ is set to be a small value, e.g., 1e-4;

3. DAG layer loss to make sure the causal representation $\mathbf{z}$ and its reconstructed self

are close to each other:

$\mathcal{L}_{causal} = ||\mathbf{z} - f(\mathbf{A} \circ \mathbf{z}; \epsilon)||_2^2$.

Thus the overall loss function is $\mathcal{L} = -\mathcal{L}_{ELBO} + \lambda_1 \mathcal{L}_{dagness} + \lambda_2 \mathcal{L}_{nonzero} + \lambda_3 \mathcal{L}_{causal}$. λ_1, λ_2, and λ_3 are the hyperparameters set to be 20, 0.1, and 1, respectively, which are selected to balance the relative value of each item in the loss function.

3.7.3 ECG Reconstruction Performance of Personalized Digital Twins

	Interpolation		Extrapolation	
	ρ	rRMSE	ρ	rRMSE
Cardiac young group				
Vanilla CVAE	0.92 (0.08)	0.38 (0.12)	0.95 (0.07)	0.29 (0.10)
Causal CVAE	**0.94 (0.04)**	**0.32 (0.10)**	**0.96 (0.04)**	**0.28 (0.10)**
Cardiac old group				
Vanilla CVAE	0.91 (0.16)	0.33 (0.22)	0.93 (0.14)	0.32 (0.18)
Causal CVAE	**0.97 (0.04)**	**0.21 (0.11)**	**0.97 (0.03)**	**0.23 (0.10)**
Noncardiac young group				
Vanilla CVAE	0.92 (0.05)	0.39 (0.11)	0.92 (0.08)	0.39 (0.13)
Causal CVAE	**0.95 (0.04)**	**0.30 (0.10)**	**0.94 (0.08)**	**0.31 (0.15)**
Noncardiac old group				
Vanilla CVAE	0.94 (0.11)	0.31 (0.16)	0.96 (0.07)	0.27 (0.12)
Causal CVAE	**0.96 (0.09)**	**0.22 (0.15)**	**0.97 (0.04)**	**0.21 (0.10)**

Table 3.7: The results from the proposed causal CVAE as the backbone model for the inferred ECG of each group in terms of the mean and the standard deviation (in parenthesis) of Pearson coefficient (ρ) and rRMSE. Transfer learning is applied by tuning the first layer of the decoder and the newly added causal layers. The vanilla CVAE comparison group is copied from Table 3.6 when tuning the first layer of the decoder for easier reference.

In this section, we examine the performance of ECG reconstruction using the proposed causal CVAE model as the backbone for transfer learning. From Chapter 3.6, we find that tuning the first layer of the decoder achieves reasonably good ECG reconstruction performance with fewer parameters to tune. Thus, for the causal CVAE model, we

also load the parameters from the leave-N-out case and fine-tune the first layer of the decoder together with the newly added causal representation learning layers. The dimension of the latent vector is chosen as 8. The ECG inference performance is listed in Table 3.7. Compared to the vanilla CVAE model results in Table 3.6, the proposed causal CVAE model achieves better results in terms of ECG reconstruction.

3.7.4 Intervention Experiment

From 3.7.1, we know that with the "do" operation intervening each of the nodes in the DAG, the children nodes change together as their parent node is changed. And the intervention can generate counterfactual outputs, indicating the underlying cause and effect represented by the corresponding nodes according to the causal system. In this section, we conduct the intervention experiment during the test time. We call the ECG reconstructed in a non-intervened way "inferred ECG", which is generated by inputting a sample from normal distribution into the causal representation learning layer and concatenating it with the PPG cycle as the new input into the decoder (Fig. 3.17). Now we intervene each of the nodes in the latent vector $\mathbf{z}$ by updating their original value (that generates the "inferred ECG") to a different value (e.g., 300) and the value of their children nodes are changed as well to form the $\tilde{\mathbf{z}}$ complying with the relation in the learned DAG adjacency matrix to further generate the "intervened ECG". Since the vector dimension is set to be 8, we analyze the impact from Node 1 to Node 8 in the intervened ECG for the target patient, in terms of both timing interval and amplitude changes. In this way, we can have a better understanding of how each of the causal representation nodes plays the role in the ECG

generation.

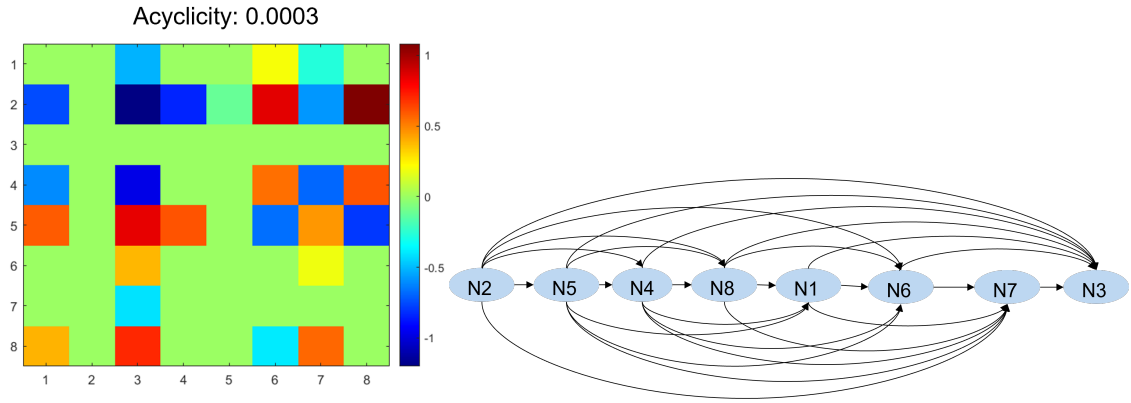
Quantitative Evaluation Metrics of Effect For Causal Analysis:

We consider the following three intervals and three wave amplitude to quantitatively evaluate the impact of the intervention, including the PR interval, the QRS duration, and the QT interval; the amplitude of the P wave, QRS complex, and T wave. **PR interval:** Normally, the PR interval lasts 0.12-0.20 seconds, which begins from the onset of the P wave and ends at the beginning of the QRS complex, representing the time for the electrical pulse to spread from the atria to ventricles through the AV node and His Bundle. We use the segment from P point to R point of ECG as the approximated PR interval. The duration of the PR interval indicates the functionality of the conduction pathway from atria to ventricles [60]. On the one hand, a prolonged PR interval can indicate the possibility of first-degree heart blockage. On the other hand, a shortened PR interval indicates either the atria have been depolarized from close to the AV node, or there is abnormally fast conduction from the atria to the ventricles. **QRS complex duration and amplitude:** The duration of the QRS complex is normally 0.12 seconds or less, for ventricular depolarization. A prolonged QRS complex indicates impaired conduction within the ventricles caused by bundle branch block or erroneous impulse pathway [60]. Increased height of the QRS complex indicates ventricular hypertrophy. **QT interval:** The QT interval is from the onset of the QRS complex to the end of the T wave, which is normally less than 0.48 seconds. An unusually prolonged or short QT interval may be due to electrolyte abnormalities or drugs [60]. **P wave amplitude:** The P wave represents the electrical activation (depolarization) of atria. If the P wave is missing or amplitude is inverted, then atria are not activated normally from the SA node. **T wave amplitude:**

The T wave shows the repolarization of the ventricles to their resting state. If the T wave is inverted, then the likely causes are ischemia or ventricular hypertrophy [60]. To summarize, in the dissertation author's understanding, the timing of the ECG represents the functionality of the impulse pathway and the shape and amplitude of the subwaves indicate the functionality of the heart muscles.

Case Study: Female, 52, Coronary Artery Disease (CAD)

We take the result of a 52-year-old female patient with CAD from the extrapolation learning mode as an example for analysis. By quantitatively evaluating the impact of the intervening nodes, we aim to infer their possible meaning in a heart process for better interpretability.



(a) (b)

Figure 3.18: (a) The learned DAG adjacency matrix $\mathbf{A}$ for a 52-year-old female subject from the cardiac young group with CAD. (b) The DAG is drawn based on the DAG adjacency matrix $\mathbf{A}$ in (a).

The visualization of the learned DAG map from the training time and the corresponding graph showing the causal relationship among different nodes in the causal latent vector $\mathbf{z}$ are illustrated in Fig. 3.18. From the DAG map in Fig. 3.18a, we know that it can be permutated in both rows and columns to form an upper triangle matrix, implying the

"DAGness" is preserved after the training with acyclicity being 0.0003. As we know in a causal graph, the intervention on a parent node will be translated to their children node, thus the fewer children a node has, the easier to analyze its independent impact on the ECG. In this case study, we focus on the impact of Nodes 3, 7, and 6 in Fig. 3.18b in our following analysis.

	Inferred ECG	Intervened ECGs							
		Node 3	Node 7	Node 6	Node 1	Node 8	Node 4	Node 5	Node 2
PR Interval (s)	0.13	0.27	0.12	0.17	0.12	0.27	0.18	0.20	0.21
QRS Complex Duration (s)	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10
QT Interval (s)	0.40	0.41	0.40	0.42	0.40	0.41	0.40	0.40	0.41
P Wave Amplitude (mV)	0.09	-0.07	0.11	0.11	0.10	-0.07	0.06	0.00	-0.01
QRS Complex Amplitude (mV)	1.15	0.73	1.03	0.89	0.96	0.69	0.92	1.01	0.68
T Wave Amplitude (mV)	0.10	0.12	0.14	0.04	0.12	0.16	0.09	0.13	0.09

Table 3.8: The mean of each evaluation metric for both inferred ECG and intervened ECGs. The results for the intervened ECGs for each node are ordered by their positions in the DAG (Fig. 3.18b).

Table 3.8 lists the averaged intervals and subwave amplitudes of the inferred ECG and the intervened ECGs. Some significant changes are concluded from the table: Tuning Node 3 increases the average PR interval length from 0.13s to 0.27s, inverts the P wave (amplitude changed from 0.09mV to -0.07mV), and reduces the amplitude of the QRS complex from 1.15 mV to 0.73mV; Tuning Node 7 (along with Node 3 because of the negative causal relation between them) leads to the 40% increase of the T wave amplitude; Tuning Node 6 (along with Node 7 and Node 3) decreases the amplitude of T wave by 60%.

In addition to the results in Table 3.8 that only show the averaged intervention effects/difference between the inferred and intervened ECGs, we plot a more detailed distribution of the difference after intervention in Fig. 3.19 to check if the impact of each node is solid. Each red circle marker represents the corresponding metric is increased

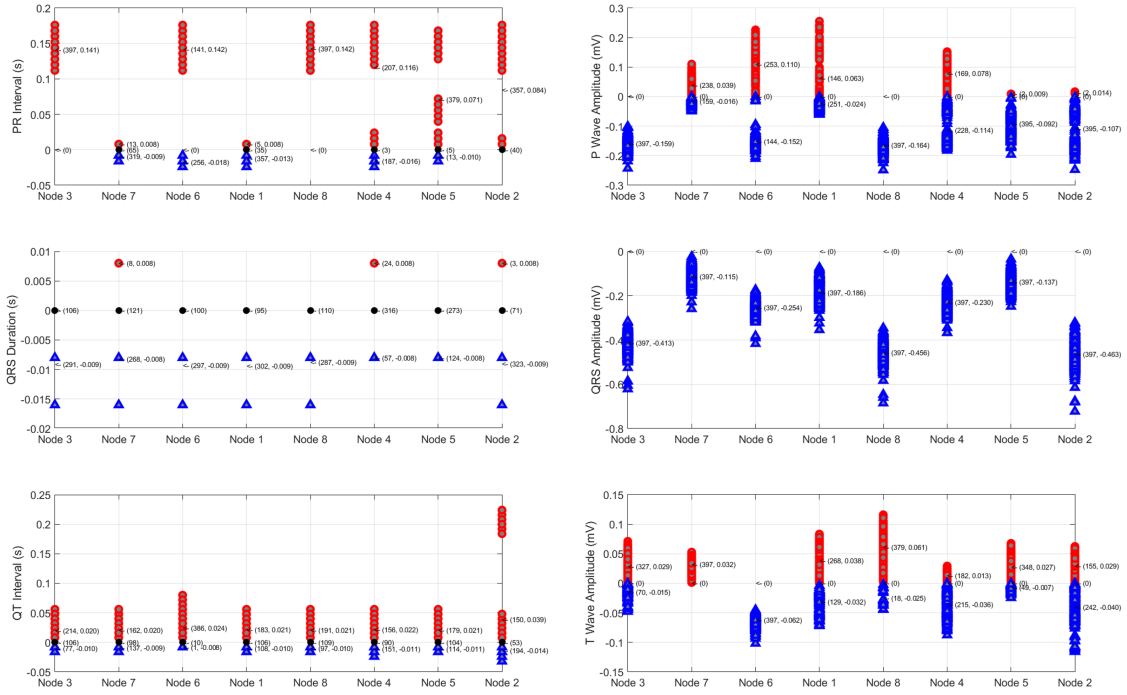


Figure 3.19: Distributions of the difference between the inferred ECG and the intervened ECGs for each evaluation metric, showing the impact of intervening each node in the latent causal representation. Red circle markers represent the increased value in the metrics after intervention and blue triangle markers represent the decreased values. For each node, the first number in each bracket represents the number of cycles that are great than, equal to, or less than zero difference after intervention and the second number represents the corresponding average of the difference.

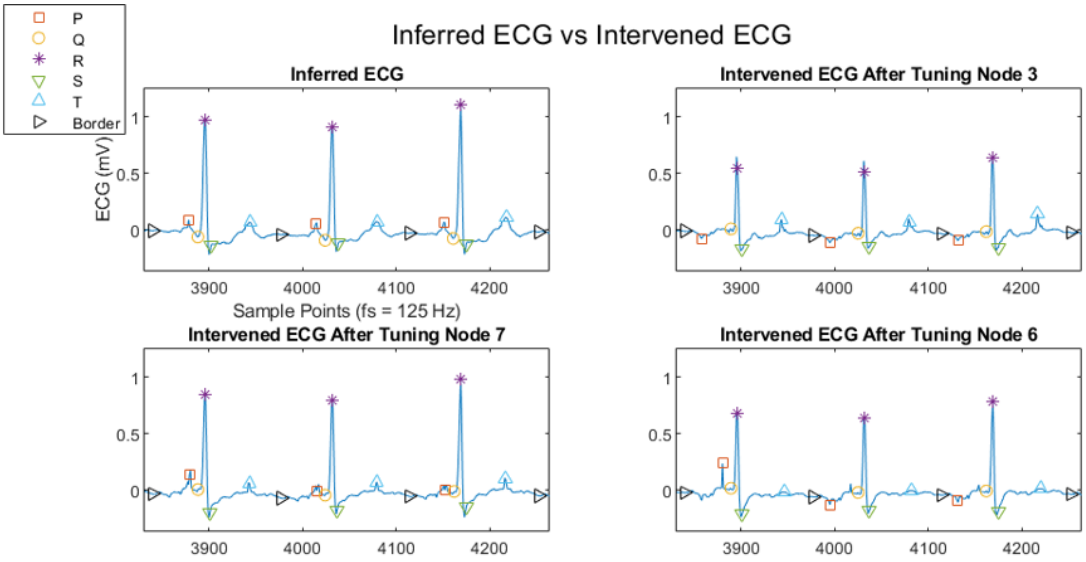


Figure 3.20: Visualization of the inferred ECG and intervened ECGs after tuning Nodes 3, 7, and 6, respectively. Established algorithms [110, 124, 125] are applied to detect the P, Q, R, S, and T fiducial points. The border point is defined as the 60%:40% segmentation point between each RR interval.

in the intervened ECG cycle than that in the inferred ECG cycle, and each blue marker represents the decreased value after intervention. For each node, there are three brackets by the side of the markers, the first number in which is the number of cycles that are greater than, equal to, or less than zero difference after intervention, respectively, and the second number in which is the corresponding average of the difference. First we examine the effects of intervening Node 3: (1) all the intervened ECG cycles have a longer PR interval than the inferred ECG with an average increase of 0.14s; (2) the P wave amplitude decreases for all ECG cycles after intervention by 0.159mV to invert the P wave; (3) all intervened ECG cycles have reduced QRS complex amplitude by an average of 0.413mV. Similarly, the effects of tuning Node 7 and Node 6, that include increasing and decreasing the T wave amplitude, are confirmed in Fig. 3.19, respectively. Note that even though tuning Node 3, 6, and 7 reduces QRS duration by approximately

0.01s and tuning Node 6 elongates the QT interval by 0.02s for almost all ECG cycles, considering the smallest time scale in the ECG grid is 0.04s, both changes are considered not significant. Also, even though the averaged P wave amplitude is shown to be increased in Table 3.8 from 0.09mV to 0.11mV after intervening Node 6 and Node 7, from Fig. 3.19, we know that this increase is not consistent across all cycles (approximately 2/3 increases and 1/3 decreases), thus this change is not considered as significant either. The intervened ECG signals generated by changing the value of Nodes 3, 7, and 6 are visualized in Fig. 3.20. From Fig. 3.20, tuning Node 3 leads to an inverted P wave, elongates the PR interval, and lowers the QRS amplitude, which are aligned with the numerical results in Table 3.8. In addition, Fig. 3.20 shows that tuning Node 7 and Node 6 leads to the peaked T wave (increased T wave amplitude) and flattened T wave (decreased T wave amplitude), which is also aligned with the numerical results in Table 3.8.

With the quantitative effect of intervening Node 3, 7, and 6 being clear, we attempt to infer what the possible physiological/medical meaning behind each of the nodes during a heart process is, i.e., trace the cause from the effect, and at the same time, their mutual causal relation should also be born in mind for inference. For example, Node 6 could be the electrolyte disorder [42, 151] that causes the lower amplitude of the T wave as shown in Fig. 3.19. And Node 7 could be the medication caused by Node 6 that helps to balance the electrolyte and leads to the increased T wave amplitude. Node 3 is the child of both Node 6 and Node 7, which could be AV block or SA block caused by the effect of drug and electrolyte abnormalities together. So that when Node 3 is intervened, prolonged PR interval impacted by the impaired conduction pathway from SA to AV node happens as the statistical results show, as well as the inverted P wave caused by improper functioning

of SA node. So far, we have examined the possibilities of the medical meaning of the latent causal representation vector using intervention with causal CVAE model. Because of the complexity of the cause of the disorders in the ECG waveform, with further professional input from doctors, our hypothesis of the interpretation can be validated and complemented for a more clinically solid causal analysis.

3.8 Chapter Summary

This chapter presents a novel application of digital twins for continuous precision cardiac monitoring by inferring ECG waveform from PPG signals. Different from the previous chapter, this chapter deals with real-world scenarios in which only limited ECG signals are available from the target individuals for whom the personalized digital twin is designed. A transfer learning method is proposed to fine-tune the generic digital twin model, which is pre-learned from a large portion of available paired PPG and ECG data from the training corpus, with limited paired PPG and ECG data from the target participant. Experimental results validate that the proposed transfer learning training scenario achieves better continuous ECG reconstruction accuracy for the target participants compared to other baseline comparison models. This suggests that our proposed method can generate a reliable digital twin for accurate and personalized continuous cardiac monitoring, providing a promising future in which people can receive early medical intervention through personalized digital twins. In addition to using the previously proposed dictionary learning framework as the backbone model for fine-tuning, the vanilla CVAE model and the causal CVAE model are proposed to learn the underlying latent vector that rep-

resents the heart process, taking the electrical and mechanical physiology process into account for better explainability and better inference performance.