DIGITAL SMART HEALTH VIA PHYSIOLOGICAL SIGNAL SENSING AND LEARNING

by

Xin Tian

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2022

Advisory Committee:
Professor Min Wu, Chair/Advisor
Professor Behtash Babadi
Professor Furong Huang
Professor Chau-Wai Wong
Professor Sushant M. Ranadive
Professor Guodong (Gordon) Gao, Dean's Representative

ABSTRACT

Title of Dissertation: Digital Smart Health Via
Physiological Signal Sensing and Learning

Xin Tian
Doctor of Philosophy, 2022

Dissertation Directed by: Professor Min Wu
Department of Electrical and Computer Engineering

Periodic blood volume change underneath a person's skin induces subtle color variations in the skin area. These subtle changes can be captured by a noninvasive and low-cost optical technique called photoplethysmography (PPG). In this dissertation, we study the modeling of contact-based and contact-free PPG signals to facilitate its promising applications in physiological signal sensing and learning for digital smart health.

In the first part of the dissertation (Ch. 2), we propose a user-friendly and continuous electrocardiogram (ECG) measurement with the help of contact-based PPG sensors for long-term cardiovascular health monitoring. ECG is a clinical gold standard for non-invasive cardiac diagnosis but continuous ECG monitoring is challenging. PPG provides a low-cost alternative, though it provides less clinical knowledge compared to ECG. How to leverage the advantages of these two measurement modalities for better and easier healthcare? We first study the physiological and signal relationship between PPG and ECG and then infer the waveform of ECG via PPG based on their relationship. Joint

dictionary learning frameworks are proposed to learn the mapping that relates the sparse domain coefficients of each PPG cycle to those of the corresponding ECG cycle. This line of research has the potential to fully utilize the easy measurability of PPG and the rich clinical knowledge of ECG for better preventive healthcare.

In the second part of the dissertation (Ch. 3), a physiological digital twin for personalized and continuous cardiac monitoring is developed. Using our proposed dictionary learning based framework as the backbone model, this part of the dissertation focuses on the problem of inferring ECG signals from PPG signals under realistic conditions where available ECG data are scarce. With transfer learning, a generic digital twin model learned from a large portion of paired ECG and PPG data is fine-tuned to precisely infer the ECG from the PPG of a target participant whose available ECG data are scarce. Experimental results validate the feasibility of using the proposed method to learn a reliable digital twin for precision continuous cardiac monitoring. Convolutional neural network based backbone model designs are also proposed based on the underlying physiological process of ECG generation for better explainability with more flexibility in transfer learning.

In the third part of the dissertation (Ch. 4 and Ch. 5), we present contactless methods of blood oxygen saturation (SpO_2) measurement from remote PPG signals captured by regular RGB smartphone cameras. Both a principled signal processing based method and a data-driven neural network based method are proposed for SpO_2 estimation by either explicitly or implicitly extracting features from multi-channel skin color signals with color channel mixing and temporal analysis. Experimental results show that our proposed methods achieve better accuracy of blood oxygen estimates compared to traditional methods using only two color channels and prior arts.

To My Parents, My Family, and Shoujing

Acknowledgments

Memory brings me back to the spring of 2018 when I was thrilled to begin my Ph.D. journey in Professor Wu's group. So many memorable years go by and I have come to this final stage of pursuing my Ph.D. degree with a more mature, humble, and grateful heart. Looking back, I owe my gratitude to all the people who have made this dissertation possible and made my graduate life a precious experience to reflect on in the future.

First of all, I'd like to express my appreciation to my advisor, Professor Min Wu, for providing me with invaluable opportunities to work on challenging and meaningful projects over the past years. Professor Wu is always supportive and has been a role model from whom I not only gain expertise but also learn to take ownership and step out of comfort zones with broader visions, curiosity, courage, and enthusiasm. It is a pleasure to have such an extraordinary and encouraging advisor, without whose professional expertise and wise guidance, this dissertation would have been a distant dream. I also want to thank Professor Behtash Babadi, Professor Furong Huang, Professor Chau-Wai Wong, Professor Sushant Ranadive, and Professor Guodong Gao for agreeing to serve on my committee and providing invaluable insights for the dissertation.

I am sincerely thankful for the group members at the MAST-UMD group who have greatly helped me: Dr. Qiang Zhu, Dr. Mingliang Chen, Mr. Zachary Lazri, Mr. Fakai Wang, Mr. Ashira Jayaweera, Ms. Yiqi Li, and Mr. Carl Steinhauser. Especially Dr.

Zhu has patiently helped me a lot when he introduced and guided me to my first research project. I would also like to acknowledge the tremendous help and insightful discussions I have had from professors and their students on multiple collaborative projects: Professor Yuenan Li, Professor Chau-Wai Wong, Mr. Joshua Mathew, Ms. Jisoo Choi, Professor Sushant Ranadive, Ms. Sara Mascone, and Ms. Emily Blake. I am also grateful to Dr. Watt for the constructive input and feedback to inspire the preliminary ideas for the digital twin project. I wish to thank all the professors at the University of Maryland who taught me courses and helped me build the stepping stones for my subsequent research.

Words cannot express the gratitude I owe to my family that has always stood by me and pulled me through seemingly impossible challenges at times. I would also like to thank Shoujing who gives me unconditional support, encouragement, care, and company during those bittersweet times in life. I dedicate this dissertation to them. I want to thank myself for having perseverance, courage, and passion to make all these good things happen on this journey.

Last but not least, I offer my regards and blessings to all who supported me in all aspects during my graduate life!

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

1.1.1 Photoplethysmography (PPG) and Remote PPG

Photoplethysmography (PPG) is a low-cost, user-friendly, and noninvasive optical technique that measures the periodic change of blood volume in the microvascular bed of tissue in the pace of heartbeat, and can be obtained from an optoelectronic device clipped to a person's fingertip or by other wearable devices such as an Apple Watch, as shown in the "contact PPG" part of Fig. 1.1. PPG has become a common modality for heart activity monitoring in clinics, hospitals, and homes for healthcare and fitness purposes [7]. As illustrated in Fig. 1.2(a), the measurement of PPG requires a light source to illuminate the tissue and a photodetector to receive the light transmitted or reflected by the tissue. During each cardiac cycle, the blood is first pumped into the body so that the blood volume increases in the capillaries in the skin, which causes increased light absorption. Then as the blood travels back to the heart via the venous network, the light absorption at the capillaries decreases. Therefore, as shown in Fig. 1.2(b) the PPG signal is composed

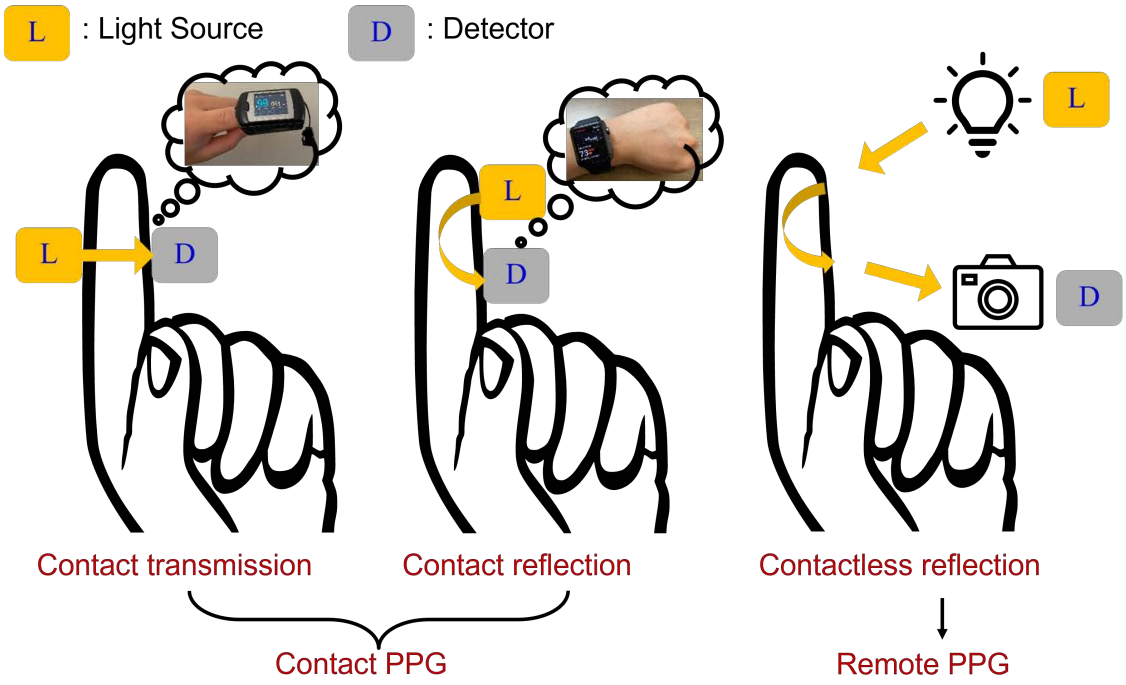


Figure 1.1: PPG measurement in both contact and contactless methods.

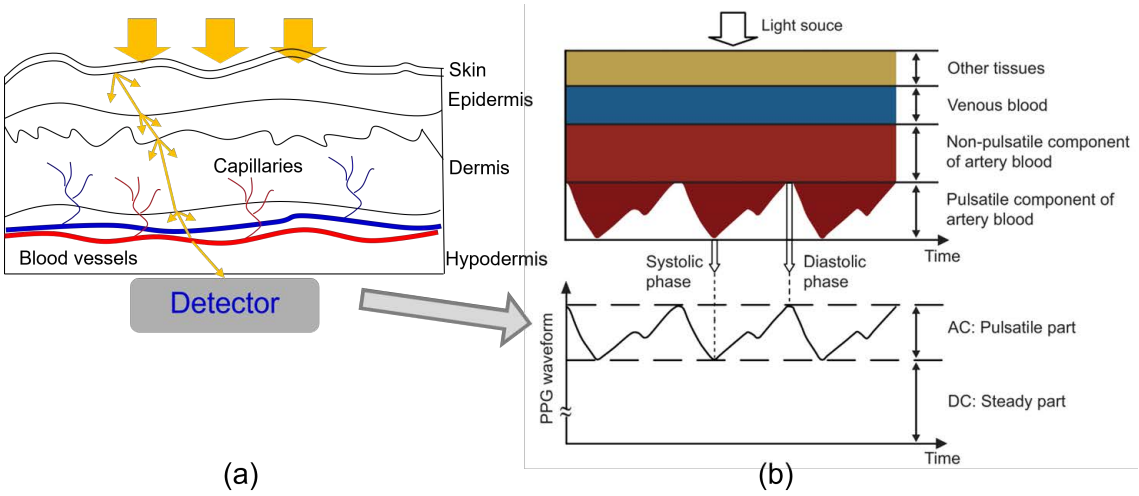


Figure 1.2: (a) Anatomical cross-section structure of human skin tissues and transmitted light captured by a detector when the skin is illuminated by a light source. (b) Variations in light attenuation by tissue, illustrating the rhythmic effect of arterial pulsation. Figures modified from [149] and [129].

of the 'AC' component which reflects the cardiac change with each heartbeat, and the 'DC' component which contains the information including respiration, venous flow, and thermoregulation [129].

To facilitate long-term and comfortable sensing, advances in video signal processing, computer vision, and artificial intelligence have opened up opportunities to use a camera captured video to monitor a person's health related vital signs remotely as illustrated in the "remote PPG" part of Fig. 1.1. This technology is commonly referred to as remote PPG (or rPPG), which is first proposed by Verkruysseet et al. [145]. The basic principle of rPPG is to illuminate tissue and to use the camera as the receiving sensor to capture the re-emitted light from the tissue. The captured video contains the periodically variational light absorption in the microvascular bed underneath the skin, thus can convey information about the cardiovascular system. The rPPG has been utilized to monitor important physiological parameters, including heart rate [34, 87, 141, 145, 150, 166], breathing rate [25, 109], heart rate variability [44, 66, 95, 109], blood pressure [68], and blood oxygen saturation [77].

In this dissertation, we study the following two application directions of contact and contactless PPG, one is electrocardiogram waveform inference from contact PPG waveform and its application and contribution to the emerging digital twin technology (Ch. 2 and Ch. 3), and the other is noncontact blood oxygen saturation measurement using remote PPG captured by RGB cameras (Ch. 4 and Ch. 5).

1.1.2 Electrocardiogram (ECG) and Its Physiological and Signal Relation With PPG

Cardiovascular diseases (CVDs) have become a leading cause of death globally. From alarming reports of the World Health Organization, an estimated 17.9 million people died from CVDs in 2019, representing 32% of all global deaths [20]. However, some CVDs, such as heart muscle dysfunction, show no obvious symptoms in the early stage. The presence of symptoms usually indicates the onset of heart failure. A study conducted on the aged population shows that around one-third to one-half of heart attacks are clinically unrecognized [35]. The unawareness of diseases makes some patients miss the opportunities of receiving an early medical intervention.

Electrocardiogram (ECG) is a widely-used clinical gold-standard for the detection of irregular heart rhythms, cardiovascular diseases, and determination of how certain heart disease treatments are working in a painless and noninvasive manner. By measuring the electrical activity of the heartbeat and conveying information regarding heart functionality, timely and continuous ECG monitoring is proven to be beneficial for the early detection of CVDs [114, 128]. The clinical standard ECG measurement device is the 12-lead ECG monitor that is commonly seen in the health care provider's office, a clinic, or a hospital room. It sensitively picks up electrical potential changes spread in the heart from the skin during a cardiac cycle from 12 different perspectives by attaching ten electrodes with sticky patches to each of the limbs and six positions across the chest of the patient.

A normal ECG waveform and the corresponding electrical and mechanical activities in the heart are shown in Fig. 1.3. Specifically, different phases in one cardiac cycle

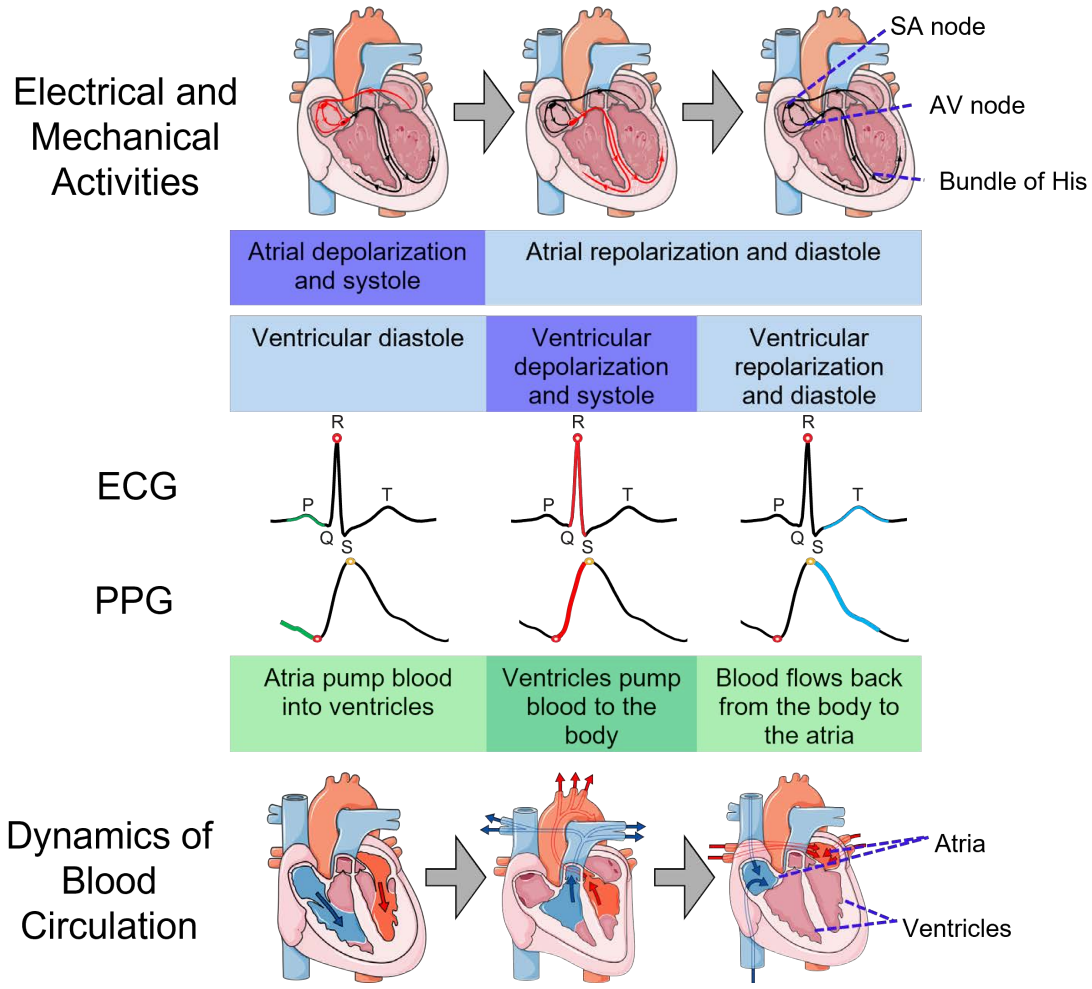


Figure 1.3: Association between the electrical and mechanical activities of the heart and the simultaneous blood flow dynamics represented by ECG and PPG, respectively. The heart images are adopted from Servier Medical Art [120].

progress as follows [8,57]: A cardiac cycle begins with the atria depolarization and systole triggered by the heart's pacemaker at the sinoatrial (SA) node, represented by the P-wave of ECG. The electrical impulse then spreads across the atria to the atrioventricular (AV) node and proceeds to the ventricular walls through the bundle of His to initiate ventricular contraction that is recorded by the QRS complex of ECG. After the ventricles are completely activated, they start to repolarize (return to the resting electrical state) and relax, and the T-wave in ECG depicts this phase. Finally, both the atria and ventricles complete repolarization and a new cycle is about to start.

As a dynamically involved system, the electrical stimulus that spreads in the heart drives the orderly contraction and relaxation of the heart muscles, leading to blood pumped into the vessels and peripheral ends that can be captured by PPG. As a result, the dynamics of blood flow are coupled with the transmission of electrical stimulus throughout the heart, indicating that PPG and ECG represent the same cardiac process measured in different signal sensing modalities. As shown in Fig. 1.3, the ascending slope of PPG is caused by ventricle contraction represented by the QRS wave complex in an ECG cycle that pumps blood to the vessels and microvasculature and increases the blood volume there correspondingly [7]. And the descending slope of PPG forms when the blood flows back from the body towards the heart during the ventricle repolarization and relaxation represented by the T-wave and the P-wave of the next cycle.

From the signal processing perspective, we view the ECG as the source signal and PPG as the signal on the receiver side through our cardiovascular system. The electrical cardiac activity caused our heart to beat, followed by the variations of the aortic blood pressure and the blood circulation all over our body, which includes the peripheral site

used to measure the PPG signal. The system from ECG to PPG can be treated as an equivalent lowpass filter given the time-frequency representation of the two signals shown in Fig. 1.4, which also inspires a way to model the relationship from PPG to ECG as an inverse engineering. In particular, the low frequency component of ECG can be recovered via some inverse filtering process from the low frequency of PPG. And the high frequency part of the ECG signal can be reconstructed by examining the correlation between the high and low frequency parts of ECG. In our previous work, the transition from PPG to ECG (route (a) + (b) in Fig. 1.4) in the frequency domain can be characterized by a linear transform relating PPG and ECG in the DCT domain [164, 165] that embodies the underlying electrical, biomechanical, and optophysiological principles.

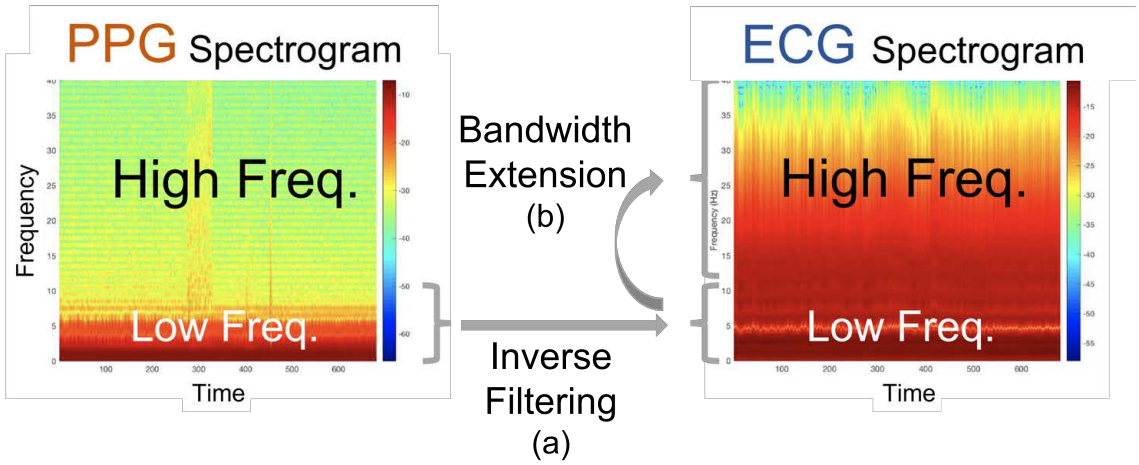


Figure 1.4: Spectrograms of PPG and ECG manifesting the route to discover and model their relationship from PPG to ECG. (a) The low frequency component of ECG can be recovered via some inverse filtering process from the low frequency of PPG. (b) The high frequency part of the ECG signal can be reconstructed by examining the correlation between the high and low frequency parts of ECG.

1.1.3 Digital Twins

Accompanying the industrial revolutions over the past several centuries have been four eras of healthcare revolutions [83, 84, 131], the most recent of which is just beginning. Technological advances have enabled improvements in patient care and monitoring by introducing portable and affordable devices such as pulse oximeters [84], as well as communication and computing infrastructures that make telehealth and remote care possible, enabling people to obtain service from the comfort of their homes [84, 131].

In addition, precision health takes various types of patient-specific information into account, enabling personalized monitoring for preventative care, early detection of diseases, and individualized treatment(s) with potentially improved outcomes.

The digital twin is a promising paradigm toward realizing precision health. As a digital representation of a physical artifact to facilitate the monitoring of the artifact's status [16], the notion of a digital twin was introduced by Michael Grieves in his 2002 presentation for product life cycle management [51, 52] and adopted by the U.S. National Aeronautic Space Administration in its aerospace missions [45, 48]. Broadening the scope of digital twins to healthcare has the potential of producing fine-grained tailor-built models of the biological phenomena relating to an individual's health [16].

Given the high anatomical complexity of human bodies, it is nearly impossible to build one digital twin model to account for all aspects of health needs. Thus, we apply a common engineering strategy of divide-and-conquer to cluster digital twins for healthcare into the following categories:

1. Organ and structural level digital twins [3, 16]: a major application of these digital

twins is to build models to help clinicians understand an individual patient's organ structure, support surgery planning, and reduce treatment risks and uncertainties;

2. Omics level digital twins [16]: these digital twins relate genome and other omics information with patients' health risks, reactions to drugs, and other high-level health conditions;

3. Physiology level digital twins [16, 91]: these digital twins leverage sensing technologies and data science to facilitate the monitoring, analysis, and management of a patient's health conditions on-demand and/or at a chosen time scale.

We focus on the physiological digital twin in this work. One representative prior work in this area [16] describes a digital twin as a system for providing fast, accurate, and efficient medical simulation, which consists of a physical object, a virtual object/model, and healthcare data [91]. A physical object can be represented as a medical or wearable device for monitoring a person's health; and the healthcare data can include data from wearable devices, real-time monitoring data, and simulation data from digital models. While increasingly large amounts of data are becoming available to potentially support data-driven approaches, we believe that healthcare digital twins research should strive for constructing explainable digital twin models when considering both the complex ethics and social aspects of healthcare.

1.1.4 Blood Oxygen Saturation

Peripheral blood oxygen saturation (SpO_2) shows the ratio of oxygenated hemoglobin to total hemoglobin in the blood, which serves as a vital health signal for the operational

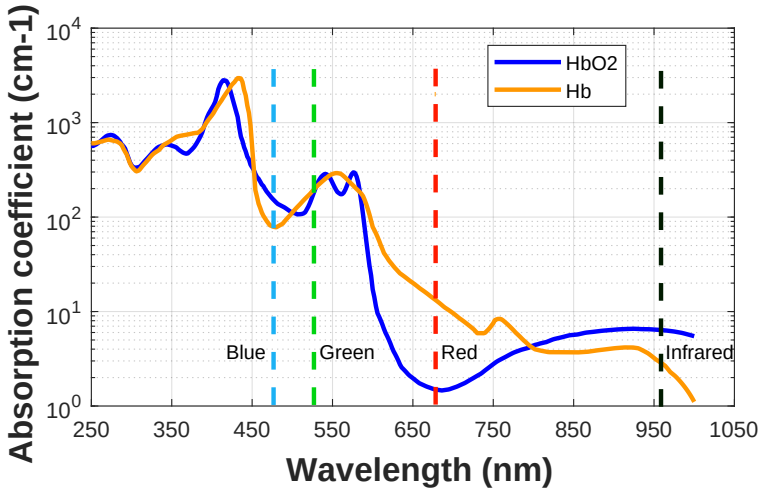


Figure 1.5: Extinction coefficient curves of hemoglobin, figure reproduced based on [37, 102]. The difference between oxygenated and deoxygenated hemoglobin at the red and blue wave lengths means these color channels contain most of the useful information for SpO_2 prediction.

functions of organs and tissues [125]. The normal range of SpO_2 is 95% to 100% [100]. Abnormality in the SpO_2 level can serve as an early warning sign of respiratory diseases [100]. The estimation and monitoring of SpO_2 are essential for the assessment of lung function and the treatment of chronic pulmonary diseases. It has become increasingly important in the COVID-19 pandemic, where many patients have experienced "silent hypoxia," a low level of SpO_2 even before obvious breathing difficulty is observed [32, 124, 138]. The vulnerable population with a high possibility of infection is recommended to monitor their oxygen status continuously for early COVID-19 detection [124, 135].

Pulse oximeters have been widely used for SpO_2 measurement at home and in hospitals in the form of a finger clip [121, 152], which adopts the *principle of ratio of ratios* (RoR) that was first proposed by Aoyagi in the early 1970s [121]. The RoR principle is based on the different optical absorption rates of the oxygenated hemoglobin (HbO_2) and

deoxygenated hemoglobin (Hb) at 660 nm (red) and 940 nm (infrared) wavelengths as indicated in Fig. 1.5. By illuminating red and infrared lights on the fingertip, the more oxygenated hemoglobin in the blood, the less infrared light and the more red light are received by the detector after transmission. In other words, the relative AC and DC amplitudes between red and infrared PPG contains pulsatile information to derive SpO_2.

The gold standard for measuring blood oxygen saturation is blood gas analysis, which is invasive and painful and requires well-trained healthcare providers to perform the test. In contrast, the pulse oximeter is noninvasive and provides readings in nearly real time, and is therefore more tolerated and convenient for daily use. The pulse oximeter is known to have a deviation of $\pm 2\%$ from the gold standard when the blood oxygen saturation is in the range of 70% to 99% [108], which is well-known and accepted in clinical use.

1.2 Main Contributions

In this dissertation, we study the modeling of contact and contactless PPG signals to facilitate its promising applications in cardiovascular signal and vital sign sensing and learning for digital health. First, we explore the potential of user-friendly and continuous electrocardiogram (ECG) monitoring with the help of fingertip PPG sensors in Ch. 2. Next, we develop a physiological digital twin for personalized continuous cardiac monitoring in Ch. 3. Last, we study the noncontact methods of blood oxygen saturation (SpO_2) monitoring from remote PPG signals captured by smartphone cameras. Both principled signal processing method and neural network based method for explicit (handcrafted) and

implicit (data-driven) feature engineering from multi-channel color signals are proposed in Ch. 4 and Ch. 5, respectively. Below are the detailed key contributions of this dissertation research.

1.2.1 Cross-domain Joint Dictionary Learning for ECG Inference from PPG

The inverse problem of inferring clinical gold-standard electrocardiogram (ECG) from photoplethysmogram (PPG) that can be measured by affordable wearable internet-of-healthcare-things (IoHT) devices is a research direction receiving growing attention. It combines the easy measurability of PPG and the rich clinical knowledge of ECG for long-term continuous cardiac monitoring. The prior art for reconstruction using a universal basis, such as discrete cosine transform (DCT), has limited fidelity for uncommon ECG waveform shapes due to the lack of representative power. To better utilize the data and improve data representation, in Ch. 2, we design two dictionary learning frameworks, the cross-domain joint dictionary learning (XDJDL) and the label-consistent XDJDL (LC-XDJDL), to further improve the ECG inference quality and enrich the PPG-based diagnosis knowledge. Building on the K-SVD technique, our proposed joint dictionary learning frameworks largely extend the expressive power by optimizing simultaneously a pair of signal dictionaries for PPG and ECG with the transforms to relate their sparse codes and disease information. The proposed models are evaluated with a variety of PPG and ECG morphologies from benchmark datasets that cover various age groups and disease types. The results show that the proposed frameworks achieve better

inference performance than previous methods, suggesting an encouraging potential for ECG screening using PPG based on the proactively learned PPG-ECG relationship. By enabling the dynamic monitoring and analysis of the health status of an individual, the proposed frameworks contribute to the emerging digital twins paradigm for personalized healthcare.

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

Digital twins are emerging as a promising framework for realizing precision health for their ability to represent an individual's health status. Ch. 3 of the dissertation work introduces a physiological digital twin for personalized and precision continuous cardiac monitoring in the form of modeling the PPG-ECG relationship. Using the dictionary learning algorithm proposed in the previous chapter as the backbone model, the work in this chapter focuses on the problem of inferring ECG signals from PPG signals for continuous precision cardiac monitoring under realistic conditions in which available ECG data is scarce. By performing transfer learning, a generic digital twin model learned from a large portion of paired ECG and PPG data is fine-tuned to precisely infer the ECG from the PPG of a target participant whose available ECG data are scarce. Experimental results for interpolation and extrapolation testing scenarios show that the proposed transfer learning method yields better ECG reconstruction accuracy compared to other baseline comparison models. This suggests that it can be used as a reliable digital twin for precision continuous cardiac monitoring. In parallel, neural network and causality based

backbone model designs are also proposed based on the underlying physiological process of ECG generation for better explainability.

1.2.3 Noncontact Hand Video Based SpO_2 Monitoring Using Smartphone Cameras

SpO_2 is an important indicator of pulmonary and respiratory functionalities. It is recommended, especially for the vulnerable population, to regularly monitor the blood oxygen level for precaution. Recent works have investigated how ubiquitous smartphone cameras can be used to infer SpO_2. Most of these works are contact-based, requiring users to cover a phone's camera and its nearby light source with a finger to capture reemitted light from the illuminated tissue. Contact-based methods may lead to skin irritation and sanitary concerns, especially during a pandemic. In this dissertation, we propose a noncontact method for SpO_2 monitoring using remote PPG signals in the hand videos acquired by smartphones. The whole algorithm pipeline includes 1) receiving video of the hand of a subject captured by a regular RGB camera of a smartphone; 2) extracting a region of interest of the hand video; 3) performing feature extraction of the region of interest based on spatial and temporal data analysis of more than two color channels; and 4) estimating a blood oxygen saturation level of the subject from the features.

The contributions of this dissertation mainly focus on the feature engineering and estimation parts of the pipeline. Considering the optical broadband nature of the red (R), green (G), and blue (B) color channels of the smartphone cameras, we exploit all three channels of RGB sensing to distill the SpO_2 information beyond the traditional ratio-of-

ratios (RoR) method that uses only two wavelengths. In the principled signal processing method (Ch. 4), the features are explicitly extracted based on the multi-channel RoR after adaptively narrow bandpass filtering based on accurately estimated heart rate to obtain the most cardiac-related AC component for each color channel. Experimental results show that our proposed blood oxygen estimation method can reach a mean absolute error of 1.26% when a pulse oximeter is used as a reference, outperforming the traditional RoR method by 25%. With the understanding of the multi-channel based principled signal processing method, convolutional neural network based schemes (Ch. 5) are further proposed for implicit data-driven feature extraction by skin color channel mixing and temporal analysis. The neural network architectures are designed inspiring by the optophysiological models for SpO_2 measurement. Through the visualization of the weights for the RGB channel combinations, we demonstrate the explainability of our model and that the choice of the color band learned by the neural network is consistent with the suggested color bands used in the optophysiological methods.

Chapter 2

Cross-domain Joint Dictionary Learning for ECG Inference from PPG

2.1 Motivation and Problem Formulation

Asymptomatic and intermittent abnormalities in the heart functionality could be missed without continuous ECG monitoring, which plays an important role in the early detection and prevention of life-threatening cardiovascular diseases. However, the conventional continuous ECG equipment (e.g., the Holter monitor for 24 to 48 hours of recording) is bulky and can be restrictive on users' activities, making it impossible to wear in a long term. Newer clinical ambulatory ECG monitoring devices, such as the Zio patch [36], are made to be light-weighted and have alleviated the above-mentioned issues, although potential skin irritation during long-term adhesive wear remains, especially for people with sensitive skin. In addition, a prescription is needed to obtain the Zio patch, thus not easily accessible to the general public. Apple Watch [132] and wearable devices alike, such as Omron KardiaMobile [73], are moderately affordable and can show real-time ECG without adhesion to the skin, but they generally require active user participation

Figure 2.1: Ambulatory and portable/wearable ECG devices (from left to right): Holter monitor, Zio Patch, KardiaMobile, and Apple Watch. Images in this figure are from [36, 63, 73, 132].

ECG Sens. Tech.	Cost	Accessibility	Need No Active Participation?	Long-Term & Conts. Monitoring
Standard ECG	High*	Low	×	×
Apple Watch	Medium	High	×	×
KardiaMobile	Low	High	×	×
Zio patch	Medium	Low	✓	✓(skin irritation)
Our Proposed	Low	High	✓	✓(little side effect)

*High cost in the U.S. if without medical insurance.

Table 2.1: Comparison of different ECG sensing techniques.

and is usually for sporadic and short measurement of 30-second periods, making it infeasible for long-term continuous ECG monitoring. Table 2.1 summarizes the comparison of different ECG sensing techniques discussed above and shown in Fig. 2.1.

Given the constraints of the ECG sensors, researchers have made efforts toward long-term continuous ECG monitoring by inferring full ECG waveform from optical sensors, such as the photoplethysmogram (PPG) sensors [136, 164, 165]. PPG sensors are ubiquitously seen in the wearable internet-of-healthcare-things (IoHT) devices and have become a common modality for monitoring heart conditions due to the maturity of the technology and low cost [55]. It measures the optical response of the blood volume changes at the peripheral ends, including fingertips [7], and provides valuable information about the cardiovascular system via daily use of the pulse oximeter. Compared to

ECG, PPG is more user-friendly in long-term continuous monitoring without constant user participation.

PPG and ECG are physiologically related as they embody the same cardiac process in two different signal sensing domains. As explained in Chapter 1.1.2, the peripheral blood volume change recorded by PPG is influenced by the contraction and relaxation of the heart muscles, which are controlled by the cardiac electrical signals triggered by the sinoatrial node [72]. The waveform shape (i.e. signal morphology), pulse interval, and amplitude characteristics of PPG provide important information about the cardiovascular system [7], including heart rate, heart rate variability [47], respiration [70], and blood pressure [30]. Therefore, inferring the medical gold-standard ECG signal using the PPG sensor provides a solution to achieve a low-cost, long-term continuous cardiac monitoring, which facilitates further diagnosis and leads to early intervention opportunities, especially for the low-income, disadvantaged populations, who have limited access to affordable preventive care. Our proposed technique embodies the trend of *digital twins* in healthcare [16], which is an emerging technology that plays a pivotal role in advancing personalized healthcare. The aspects of digital twins that our work contributes to are on developing a rich representation of an individual supported by data and models, through which the physiological status of this individual can be dynamically monitored and analyzed over time.

2.2 Related Works

2.2.1 ECG reconstruction from PPG

There are many prior arts extracting physiological parameters [13, 159] or classifying arrhythmia [1, 14, 58, 105] from the input ECG or PPG signals using machine learning methods. However, direct parameter estimation or automatic diagnosis is insufficient for medical practitioners to interpret. The ECG signal, rather than the derived results via black-box models, is still the gold-standard tool on which cardiologists rely and make further decisions. Our proposed technique in this chapter providing the reconstructed ECG waveform offers complementary support and allows the manual check from cardiovascular experts with their medical expertise and clinical experiences.

Very limited prior work has been devoted to the PPG-based ECG inference. The pilot study [164, 165] proposed to relate the waveforms of PPG and ECG in the discrete cosine transform (DCT) domain by a linear model. In the participant-specific case where a linear model is trained from and tested for the same individual, this DCT method achieved a mean reconstruction correlation of 0.94. In contrast, for the group-based model, the achieved mean correlation degraded to 0.79. This suggests that there is still substantial room for improvement when extending to the group-based model case where a universal mapping needs to be trained by a wider variety of ECG morphologies from multiple people. To address these above-mentioned issues, we consider dictionary learning based sparse representation for ECG and PPG as it provides a richer and more adaptive representation than the universal dictionary DCT by better leveraging data. And we will use

this as a foundation to develop joint dictionary learning models for reconstruction.

2.2.2 Dictionary learning

Algorithms that learn a single dictionary for signal representation have been well-studied [2, 41, 93]. They have been successfully applied to cardiac signal processing, including recent research showing that ECG signals can be well-represented as a sparse linear combination of atoms from an appropriately learned dictionary for such applications as ECG classification and compression [33, 90, 94].

In the domain of image processing and computer vision, these single dictionary learning strategies have been extended to joint dictionary learning tasks. For image super-resolution [154–156], coupled dictionary learning frameworks are proposed to learn a dictionary pair for low- and high-resolution image patches while enforcing the similarity of their sparse codes with respect to their dictionaries. One assumption from this model is that the transform matrix between the two sparse codes is an identity matrix. In person re-identification [85] and photo-to-sketch [148] problems, a linear mapping between the codings of input and output images is introduced into the objective function for semi-coupled dictionary learning. In both training schemes, the updates of the mapping and dictionaries are separately done within each iteration, making the dictionary computation less aware of the signal transform.

Our method aims at boosting reconstruction performance from PPG to ECG by using a joint dictionary learning framework. Unlike the super-resolution problem [154–156] where the input and output reside in the same signal domain and are highly correlated,

XDJDL introduces a PPG-to-ECG mapping, which spans the two sensing modalities with low waveform correlation, providing more flexibility and generalization for the two learned dictionaries. Different from [85, 148], we update the linear transform and the dictionary in the same step, which can optimize the capability of the obtained dictionaries for both signal representation and transformation. This kind of transform-aware joint dictionary learning formulation is one of the major differences from other coupled dictionary learning frameworks. This framework can also be easily generalized to different constraints. For instance, in the proposed LC-XDJDL model, we add a label-consistency regularization term to the objective function of the XDJDL model, which encourages the transformed sparse codes from the same class to be similar.

2.3 Proposed Methods

The previous work of ECG reconstruction from PPG using a universal, data-independent basis of the discrete cosine transform (DCT) [164, 165] has limited fidelity to represent uncommon ECG waveform shapes, especially for the *group-based* case with a broader range of signal morphologies [155]. We focus on such group-based cases in this chapter and consider data science and learning techniques with richer representative power to answer the following research question:

- *Group-based model*: Can a single model, trained from a group of subjects with a certain determinant of physiology (e.g., age, weight, disease type, etc.), predict the ECG waveforms from unseen PPG measurements for individuals in the training group?

To overcome the limitation of the DCT method and develop the synergy of model and data, our work aims at improving data representation through a more versatile and adaptive framework based on dictionary learning to demonstrate the feasibility of ECG waveform inference from PPG signal as an inverse filtering problem. In addition to the algorithmic improvement, sparse coding and dictionary learning frameworks are proven to perform efficiently in IoT platforms in terms of cutting down power consumption and computation cost [6,89]. Thus, by investigating the dictionary learning based approach in this chapter, we strike a balance between the model complexity and practical cost in IoT applications.

Our proposed cross-domain joint dictionary learning (XDJDL) method for ECG reconstruction from PPG is summarized in Fig. 2.2. A further-developed label-consistent XDJDL model (LC-XDJDL) is also proposed when the label information for the ECG/PPG cycles is available. The PPG and ECG signals are first preprocessed into normalized signal cycles to facilitate the subsequent training. In the training phase, the ECG/PPG dictionary pair is jointly updated with a stable linear mapping that relates the sparse representations of the two measurements. In LC-XDJDL, one more linear mapping that enforces the label consistency for the PPG sparse codes will be learned to further improve the ECG reconstruction performance and enrich the PPG diagnosis knowledge base.

2.3.1 Signal Preprocessing

To establish the quantitative relationship between the corresponding cycles of ECG and PPG, we preprocess the two signals during the training phase to obtain temporally

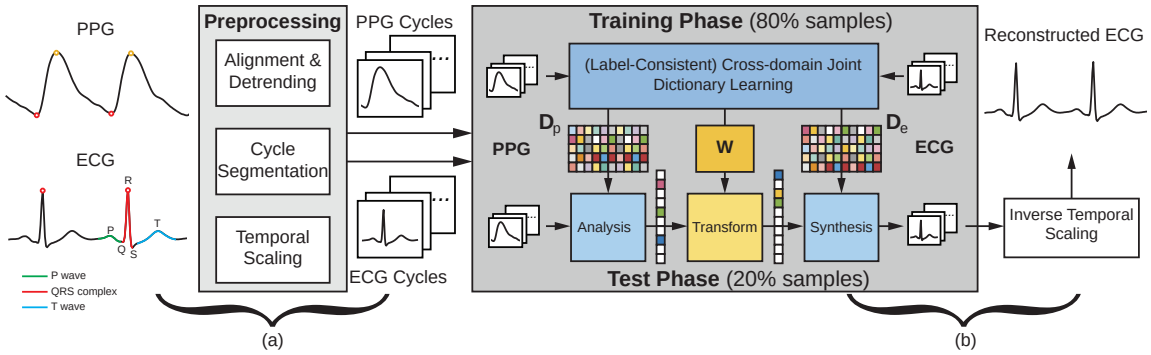


Figure 2.2: Illustration of the proposed framework. The ECG and PPG signals are first preprocessed to obtain temporally aligned and normalized pairs of cycles. 80% pairs of ECG and PPG signal cycles from each subject are used for training paired dictionaries $\mathbf{D}_p$, $\mathbf{D}_e$, and a linear transform $\mathbf{W}$ which will be applied in the test phase to infer the ECG signals.

aligned and normalized pairs of signals, so that the critical temporal features of both waveforms are synchronized for learning and evaluation. The preprocessing method we adopt is rooted in the aforementioned underlying physiological relationships between PPG and ECG signals in Chapter 1.1.2, which is independent of the dataset selection. First, considering the synchronization issue between separate ECG and PPG devices, we align the whole ECG and PPG sequences according to the moment when the ventricles contract and the blood flows to the vessels, which corresponds to the R peaks of ECG and the onsets of PPG in the same cycle. Both the onset and R peaks are detected by the beat detection functions from the PhysioNet Cardiovascular Signal Toolbox [146]. Then we detrend the aligned signals by a second-order difference operator based algorithm [164] to eliminate the baseline drift related to respiration, motion, vasomotor activity, and change in contact surface [7]. To prepare for the learning of the cycle-wise relation during one heartbeat, the detrended PPG and ECG signals are partitioned into cycles by the *R2R* [164] segmentation scheme, where R-peaks of the concurrent ECG waveform are used to partition the

signals on a heartbeat-by-heartbeat basis. After the segmentation, each cycle is linearly interpolated to length d to mitigate the influence of the heart rate variation. Finally, we normalize the amplitude of each cycle by subtracting the sample mean and dividing by the sample standard deviation. The preprocessed PPG and ECG signal cycles are stored in data matrices $\mathbf{P}$ and $\mathbf{E}$, respectively.

2.3.2 Cross-domain Joint Dictionary Learning (XDJDL)

We denote the PPG and ECG datasets as $\mathbf{P} = [\mathbf{X}_p, \mathbf{T}_p] \in \mathbb{R}^{d\times(n+m)}$ and $\mathbf{E} = [\mathbf{X}_e, \mathbf{T}_e] \in \mathbb{R}^{d\times(n+m)}$ respectively. Each column of $\mathbf{P}$ and $\mathbf{E}$ 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 cycle during the same cardiac cycle. The goal is to learn the patterns (in terms of dictionaries, mappings, etc.) from the training data $\mathbf{X}_p \in \mathbb{R}^{d\times n}$ and $\mathbf{X}_e \in \mathbb{R}^{d\times n}$ to infer the test ECG dataset $\mathbf{T}_e \in \mathbb{R}^{d\times m}$ from PPG $\mathbf{T}_p \in \mathbb{R}^{d\times m}$.

We formulate our XDJDL framework as:

$$\begin{aligned} \min_{\mathbf{D}_e,\mathbf{A}_e,\mathbf{D}_p,\mathbf{A}_p,\mathbf{W}} & \|\mathbf{X}_e - \mathbf{D}_e\mathbf{A}_e\|_F^2 + \alpha\|\mathbf{X}_p - \mathbf{D}_p\mathbf{A}_p\|_F^2 + \beta\|\mathbf{A}_e - \mathbf{W}\mathbf{A}_p\|_F^2 \\ \text{s.t. } & \|\mathbf{a}_{p,j}\|_0 \le t_p, \text{and } \|\mathbf{a}_{e,j}\|_0 \le t_e, \quad j = 1, ..., n. \end{aligned} \tag{2.1}$$

where $\mathbf{D}_p \in \mathbb{R}^{d\times k_p}$ and $\mathbf{D}_e \in \mathbb{R}^{d\times k_e}$ are dictionaries learned for $\mathbf{X}_p$ and $\mathbf{X}_e$, respectively; $\mathbf{A}_p \in \mathbb{R}^{k_p\times n}$ and $\mathbf{A}_e \in \mathbb{R}^{k_e\times n}$ are the corresponding sparse coding matrices related with the data matrices $\mathbf{X}_p, \mathbf{X}_e$ when $\mathbf{D}_p, \mathbf{D}_e$ are the current dictionaries. Each column of $\mathbf{A}_p$ and $\mathbf{A}_e$ is denoted as $\mathbf{a}_{p,j}$ and $\mathbf{a}_{e,j}$ with the sparsity upper bounded by t_p and t_e.

For the objective function in Eq. (2.1), $\|\mathbf{X}_e - \mathbf{D}_e\mathbf{A}_e\|_F^2$ and $\|\mathbf{X}_p - \mathbf{D}_p\mathbf{A}_p\|_F^2$ are the

data fidelity terms for ECG and PPG cycle sets, respectively. The term $\|\mathbf{A}_e - \mathbf{W}\mathbf{A}_p\|_F^2$ represents the mapping error between the sparse coding coefficients of ECG and PPG signals, which enforces the transformed sparse codes of PPG to approximate that of ECG. Intuitively, we can enforce the two sparse representations for ECG and PPG from the same cycle to be the same and set the regularization term as $\|\mathbf{A}_e - \mathbf{A}_p\|_F^2$. However, since ECG and PPG are from two different signal sensing modalities and the waveform difference between the two signals is significant, directly pushing their sparse representations to be similar could compromise the generalization of the two learned dictionaries.

From the formulation in Eq. (2.1), we can jointly learn the dictionaries for ECG and PPG datasets, which produce a good representation for each sample in the training set with strict sparsity constraints. Meanwhile, we learn the linear approximation of the transform that relates the sparse codes of PPG and ECG and use it to entail the intrinsic relationship between certain PPG atoms and ECG atoms from their corresponding dictionaries.

The optimization process is described as follows. Eq. (2.1) can be rewritten as:

$$\min_{\substack{\mathbf{D}_e,\mathbf{A}_e,\mathbf{D}_p,\\ \mathbf{A}_p,\mathbf{W}}} \left\| \begin{pmatrix} \mathbf{X}_e \\ \sqrt{\alpha}\mathbf{X}_p \\ \mathbf{0} \end{pmatrix} - \begin{pmatrix} \mathbf{D}_e & \mathbf{0} \\ \mathbf{0} & \sqrt{\alpha}\mathbf{D}_p \\ -\sqrt{\beta}\mathbf{I} & \sqrt{\beta}\mathbf{W} \end{pmatrix} \begin{pmatrix} \mathbf{A}_e \\ \mathbf{A}_p \end{pmatrix} \right\|_F^2 \tag{2.2}$$
$$\text{s.t. } \|\mathbf{a}_{e,j}\|_0 \leq t_e, \text{and } \|\mathbf{a}_{p,j}\|_0 \leq t_p, \quad j = 1, ..., n.$$

where $\mathbf{I}$ is an identity matrix and $\mathbf{0}$ is a zero matrix, with valid dimensions for matrix multiplication.

Let $\mathbf{X} \triangleq (\mathbf{X}_e, \sqrt{\alpha}\mathbf{X}_p, \mathbf{0})^{\mathrm{T}} \in \mathbb{R}^{(2d+k_e)\times n}$, $\mathbf{D} \triangleq (\mathbf{D}_e, \mathbf{0}, -\sqrt{\beta}\mathbf{I}; \mathbf{0}, \sqrt{\alpha}\mathbf{D}_p, \sqrt{\beta}\mathbf{W})^{\mathrm{T}} \in$

$\mathbb{R}^{(2d+k_e)\times(k_e+k_p)}$, and $\mathbf{A} \triangleq (\mathbf{A}_e, \mathbf{A}_p)^{\mathrm{T}} \in \mathbb{R}^{(k_e+k_p)\times n}$. The optimization of (2.2) can be written as the following problem:

$$\begin{aligned} &\min_{\mathbf{D},\mathbf{A}} \|\mathbf{X} - \mathbf{D}\mathbf{A}\|_F^2, \\ &\text{s.t. } \|\mathbf{a}_{+,j}\|_0 \leq t_e, \text{and } \|\mathbf{a}_{-,j}\|_0 \leq t_p, \quad j = 1, ..., n. \end{aligned} \tag{2.3}$$

where $\mathbf{a}_{*,j}$ represents the column of $\mathbf{A}_*$, and $\mathbf{A}_+$ is defined as the first k_e rows of sparse matrix $\mathbf{A}$ while $\mathbf{A}_-$ is the last k_p rows of sparse matrix $\mathbf{A}$. The formulation in Eq. (2.3) is now similar to the original K-SVD formulation [2], suggesting that K-SVD can be adapted for this optimization. The difference is the local sparsity constraint, which will be addressed in the following optimization procedures.

Step 0: Initialization

To initialize $\mathbf{D}$ and $\mathbf{A}$, we need to initialize their components: $\mathbf{D}_e, \mathbf{D}_p, \mathbf{W}, \mathbf{A}_e$, and $\mathbf{A}_p$. First, we randomly select a subset of columns from training data $\mathbf{X}_e$ and $\mathbf{X}_p$ to form $\mathbf{D}_e$ and $\mathbf{D}_p$. Then, we initialize the sparse codes $\mathbf{A}_e$ and $\mathbf{A}_p$ by solving Eq. (2.6) with respect to $\{\mathbf{D}_e,\ \mathbf{X}_e,\ t_e\}$ and $\{\mathbf{D}_p,\ \mathbf{X}_p,\ t_p\}$, respectively. Finally, we use the ridge regression model to initialize $\mathbf{W}$:

$$\min_{\mathbf{W}} \|\mathbf{A}_e - \mathbf{W}\mathbf{A}_p\|_F^2 + \lambda \|\mathbf{W}\|_F^2. \tag{2.4}$$

This has a closed-form solution as:

$$\mathbf{W} = \mathbf{A}_e\mathbf{A}_p^{\mathrm{T}}(\mathbf{A}_p\mathbf{A}_p^{\mathrm{T}} + \lambda\mathbf{I})^{-1}. \tag{2.5}$$

After the initialization, we use a two-step iterative optimization to minimize the energy in (2.3), whereby step one is sparse coding and step two is dictionary updating by SVD.

Step 1: Sparse coding

Given D, the step of sparse coding finds the sparse representation $\mathbf{a}_j$ for $\mathbf{x}_j$, for $j = 1, ..., n$, by solving

$$\min_{\mathbf{a}_j} \|\mathbf{x}_j - \mathbf{D}\mathbf{a}_j\|_2^2 \qquad \text{s.t.} \quad \|\mathbf{a}_j\|_0 \leq t. \tag{2.6}$$

where $\mathbf{a}_j$ is the j^{th} column of the sparse representation matrix $\mathbf{A}$ and $\mathbf{x}_j$ is the j^{th} training sample in matrix $\mathbf{X}$.

Many approaches were proposed to solve Eq. (2.6) [160]. Here we adopt the orthogonal matching pursuit (OMP) method [139], which is a greedy method that provides a good approximation. As mentioned earlier, the local sparsity constraints imposed on Eq. (2.3) will affect the direct application of OMP. One workaround is to solve the following problem in Eq. (2.7) in place of Eq. (2.3),

$$\min_{\mathbf{D},\mathbf{A}} \|\mathbf{X} - \mathbf{D}\mathbf{A}\|_F^2 \qquad \text{s.t.} \quad \|\mathbf{a}_j\|_0 \leq t_e + t_p, \quad j = 1, ..., n. \tag{2.7}$$

where $\mathbf{a}_j$ is the vertical concatenation of $\mathbf{a}_{+,j}$ and $\mathbf{a}_{-,j}$ in Eq. (2.3), and t_e and t_p are the sparsity constraints for the upper and bottom parts of $\mathbf{a}_j$, respectively. During the OMP process in each iteration, we will only keep the largest sparse coefficients in $\mathbf{a}_j$ to ensure

the local sparsity constraints.

Step 2: Dictionary update

To update the k^{th} atom, $\mathbf{d}_k$, in dictionary $\mathbf{D}$ and its corresponding coefficients, $\mathbf{a}_R^k$, in the k^{th} row of $\mathbf{A}$, we apply SVD to the residue term $\mathbf{R}_k \triangleq \mathbf{X} - \sum_{j \neq k} \mathbf{d}_j \mathbf{a}_R^j$. In practice, we only select the training samples that use the atom $\mathbf{d}_k$ and avoid filling in the zeros entries of $\mathbf{a}_R^k$ during the update. We do so by denoting the nonzero entries in $\mathbf{a}_R^k$ as $\tilde{\mathbf{a}}_R^k$, and correspondingly, $\mathbf{R}_k$ as $\tilde{\mathbf{R}}_k$. The updated atom $\mathbf{d}_k$ and the related coefficients $\tilde{\mathbf{a}}_R^k$ will then be computed by:

$$\min_{\mathbf{d}_k, \tilde{\mathbf{a}}_R^k} \left\| \tilde{\mathbf{R}}_k - \mathbf{d}_k \tilde{\mathbf{a}}_R^k \right\|_F^2. \tag{2.8}$$

To solve Eq. (2.8), we use the SVD method on the residue term [2], i.e. $\tilde{\mathbf{R}}_k = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathbf{T}}$. And then, $\mathbf{d}_k$ and $\tilde{\mathbf{a}}_R^k$ can be updated as follows:

$$\mathbf{d}_k = \mathbf{U}(:,1), \quad \tilde{\mathbf{a}}_R^k = \mathbf{\Sigma}(1,1)\mathbf{V}^{\mathbf{T}}(1,:). \tag{2.9}$$

Note that taking $\mathbf{D} \triangleq (\mathbf{D}_e, \mathbf{0}, -\sqrt{\beta}\mathbf{I}; \mathbf{0}, \sqrt{\alpha}\mathbf{D}_p, \sqrt{\beta}\mathbf{W})^{\mathrm{T}}$ as a whole in the dictionary update phase does not solve this optimization problem because the zero matrices part and the identity matrix part in $\mathbf{D}$ cannot be guaranteed in the update of the dictionary by SVD. A remedy to the above problem is to decompose the dictionary update problem for $\mathbf{D}$ into the following two subproblems by revisiting the matrix form of the optimization problem in Eq. (2.2).

(i) Update $\mathbf{D}_e$, $\mathbf{A}_e$:

$$< \mathbf{D}_e^*, \mathbf{A}_e^* >= \underset{\mathbf{D}_e, \mathbf{A}_e}{\operatorname{argmin}} \|\mathbf{X}_e - \mathbf{D}_e \mathbf{A}_e\|_F^2 . \tag{2.10}$$

We use SVD to update all atoms in $\mathbf{D}_e$ and the corresponding nonzero entries in $\mathbf{A}_e$ by solving Eq. (2.10) with the same procedure as in Eq. (2.8) and (2.9). The columns of $\mathbf{D}_e$ are l_2 normalized.

(ii) Update $\mathbf{D}_p$, $\mathbf{A}_p$, and $\mathbf{W}$:

The updated ECG sparse representation matrix $\mathbf{A}_e^*$ from the subproblem (i) then serves as an input to the second subproblem here to update $\mathbf{W}$, $\mathbf{D}_p$, and $\mathbf{A}_p$ in Eq. (2.11).

$$< \mathbf{D}_p^*, \mathbf{A}_p^*, \mathbf{W}^* >= \underset{\mathbf{D}_p, \mathbf{A}_p, \mathbf{W}}{\operatorname{argmin}} \left\| \begin{pmatrix} \sqrt{\alpha}\mathbf{X}_p \\ \sqrt{\beta}\mathbf{A}_e^* \end{pmatrix} - \begin{pmatrix} \sqrt{\alpha}\mathbf{D}_p \\ \sqrt{\beta}\mathbf{W} \end{pmatrix} \mathbf{A}_p \right\|_F^2 . \tag{2.11}$$

We treat $(\sqrt{\alpha}\mathbf{D}_p, \sqrt{\beta}\mathbf{W})^{\mathrm{T}}$ as a whole dictionary, and use the SVD method in Eq. (2.8) and (2.9) to update it together with the nonzero entries in $\mathbf{A}_p$. The linear transform and the dictionary are updated simultaneously, which addresses the problem of isolated update raised in [85,148] and is one of the major differences from other coupled dictionary learning models. After solving the two subproblems, $\mathbf{D}$ and $\mathbf{A}$ can be assembled by filling in the submatrices. The main steps of XDJDL are summarized in Algorithm 1.

2.3.3 Label Consistent XDJDL (LC-XDJDL)

For cases where the disease type is known or can be predicted, such as from the PPG signals that we have, we can further leverage the disease label. In this section,

Algorithm 1 Cross-domain joint dictionary learning

Input: Training data $\mathbf{X}_e$ and $\mathbf{X}_p$ of ECG and PPG cycles, Testing data $\mathbf{T}_e$ and $\mathbf{T}_p$, and sparsity constraints $t_e,\ t_p$

Training phase:
Initialization:

- Initialize $\{\mathbf{D}_e, \mathbf{D}_p\}$ by randomly selecting atoms from the training data.
- Initialize $\mathbf{A}_e, \mathbf{A}_p$ by solving Eq. (2.6) with OMP.
- Initialize $\mathbf{W}$ by Eq. (2.5).

while not converged **do**

- Update $\mathbf{D}, \mathbf{A}$ by combining updated submatrices.
- Sparse coding: compute $\mathbf{A}$ in Eq. (2.6) with OMP. Zero out the smallest nonzero entries in the columns of $\mathbf{A}$ if any local sparsity constraint does not hold.
- Dictionary update:
  - Update $\mathbf{D}_e, \mathbf{A}_e$ in Eq. (2.10) by the SVD method illustrated in Eq. (2.8)(2.9).
  - Update $\mathbf{D}_p, \mathbf{A}_p, \mathbf{W}$ in Eq. (2.11) by the SVD method illustrated in Eq. (2.8)(2.9).

end while
Testing phase:
for each sample $\mathbf{t}_p^j \in \mathbf{T}_p$ **do**

- Compute sparse code $\mathbf{s}_p^j$ of $\mathbf{t}_p^j$ under $\mathbf{D}_p$ using Eq. (2.6).
- Calculate $\mathbf{s}_e^j = \mathbf{W}\mathbf{s}_p^j$.
- Compute the reconstructed ECG sample as $\mathbf{r}_e^j = \mathbf{D}_e\mathbf{s}_e^j$, and store it in matrix $\mathbf{R}_e$.

end for

Output: $\mathbf{R}_e$

we examine the effect of adding a label consistency regularization term to the objective function in Eq. (2.1) as follows:

$$\min_{\mathbf{D}_e,\mathbf{A}_e,\mathbf{D}_p,\mathbf{A}_p,\mathbf{W}} \|\mathbf{X}_e - \mathbf{D}_e\mathbf{A}_e\|_F^2 + \alpha \|\mathbf{X}_p - \mathbf{D}_p\mathbf{A}_p\|_F^2 + \beta \|\mathbf{A}_e - \mathbf{W}\mathbf{A}_p\|_F^2 + \gamma \|\mathbf{Q} - \mathbf{H}\mathbf{A}_p\|_F^2$$
$$\text{s.t. } \|\mathbf{a}_{p,j}\|_0 \leq t_p, \text{and } \|\mathbf{a}_{e,j}\|_0 \leq t_e, \quad j = 1, ..., n. \tag{2.12}$$

where $\mathbf{Q} \triangleq [q_1, q_2, ..., q_n] \in \mathbb{R}^{r\times n}$ is a discriminative representation matrix [69] in which each column $q_i = [0, 0, ..., 0, 1, 1, 0, .., 0]^{\mathrm{T}} \in \mathbb{R}^{r\times 1}$ corresponds to a discriminative coding for an input signal. The nonzero elements in q_i occur at the corresponding disease label, which is similar to the one-hot encoding with the number of ones as a tunable parameter. The additional regularization term $\|\mathbf{Q} - \mathbf{H}\mathbf{A}_p\|_F^2$ represents the discriminative sparse code error, which enforces the transformed sparse codes of PPG to approximate the discriminative codes in $\mathbf{Q}$. It yields such dictionaries that the signals from the same class have very similar sparse codes, i.e. enforcing the label consistency in the sparse representations.

We add the label-consistency regularization term for two main purposes: One is to improve the ECG reconstruction quality by using additional class information to constrain the degrees of freedom of the PPG sparse codes. The other is to enrich the knowledge base of PPG for the diagnosis of a certain set of diseases of interest. CVDs weaken the heart functionality, which further impacts the blood circulation in the body, thus PPG manifests certain disease information. By enforcing the consistency between the sparse codes of PPG and disease labels, one can gain insights into how the disease is revealed on PPG by inspecting the specific columns of the PPG sparse coding matrix $\mathbf{A}_p$ and the

label matrix $\mathbf{Q}$.

Similarly, Eq. (2.12) can be written in the matrix form:

$$\min_{\substack{\mathbf{D}_e,\mathbf{A}_e,\mathbf{D}_p,\\ \mathbf{A}_p,\mathbf{W},\mathbf{H}}} \left\| \begin{pmatrix} \mathbf{X}_e \\ \sqrt{\alpha}\mathbf{X}_p \\ \mathbf{0} \\ \sqrt{\gamma}\mathbf{Q} \end{pmatrix} - \begin{pmatrix} \mathbf{D}_e & \mathbf{0} \\ \mathbf{0} & \sqrt{\alpha}\mathbf{D}_p \\ -\sqrt{\beta}\mathbf{I} & \sqrt{\beta}\mathbf{W} \\ 0 & \sqrt{\gamma}\mathbf{H} \end{pmatrix} \begin{pmatrix} \mathbf{A}_e \\ \mathbf{A}_p \end{pmatrix} \right\|_F^2 \tag{2.13}$$

$$\text{s.t. } \|\mathbf{a}_{e,j}\|_0 \leq t_e, \text{and } \|\mathbf{a}_{p,j}\|_0 \leq t_p, \quad j = 1, ..., n.$$

The two-step optimization method in Chapter 2.3.2 can still be applied to find the optimal solution to both the dictionary pair and the linear mappings $\mathbf{W}$ and $\mathbf{H}$. In the test phase, the PPG sparse representation matrix $\mathbf{A}_p$ is obtained by applying sparse coding with the learned $\mathbf{D}_p$, $\mathbf{H}$, the test sample matrix $\mathbf{T}_p$, and the label matrix $\mathbf{Q}$.

2.4 Experimental Evaluation

2.4.1 Dataset

The Medical Information Mart for Intensive Care III (MIMIC-III) [49, 71] is a publicly-available database assembled by researchers at MIT. It comprises a large number of ICU patients with de-identified health data from their hospital stays. To evaluate our proposed framework and algorithm, we have extracted a small subset of the MIMIC-III database as follows. First, we select waveforms that contain both lead-II ECG and PPG signals sampled at 125Hz from the MIMIC-III waveform database. Then the se-

Cardiovascular Diseases		# of Patients	# of Cycles
Congestive Heart Failure (CHF)		7	7075 (20.6 %)
Myocardial Infarction (MI)	ST-Segment Elevated (STEMI)	3	2962 (8.7 %)
	Non-ST Segment Elevated (NSTEMI)	4	4144 (12.1 %)
Hypotension (HYPO)		7	8281 (24.2 %)
Coronary Artery Disease (CAD)		12	11781 (34.4 %)
Total		33	34243 (100 %)

Table 2.2: Composition of the collected mini-MIMIC-33 dataset.

lected waveforms are cross-referenced with the corresponding patient profile by subject ID in the MIMIC-III clinical information database. Patients with the four types of CVDs are further selected: congestive heart failure (CHF), myocardial infarction (MI) including ST-segment elevated (STEMI) and non-ST segment elevated (NSTEMI), hypotension (HYPO), and coronary artery disease (CAD). These diseases are all included in the "diseases of the circulatory system" in the ICD-9 international disease classification codes. After that, we analyze the signal pair quality using the PPG SQI function from the PhysioNet cardiovascular signal toolbox [146] and keep the pair segments that are evaluated as "acceptable" or "excellent."

The resulting mini-MIMIC-33 dataset consists of 33 patients, with each patient having only one of the four diseases in the record. Each patient has three sessions of 5-min ECG and PPG paired recordings collected within several hours, resulting in 34243 ECG/PPG cycle pairs in total. Table 2.2 shows the composition of the collected dataset.

2.4.2 Metrics for Evaluation

As shown in Fig. 2.4 (a), a complete ECG cycle contains five major points, including P, Q, R, S, and T, which segment the ECG cycle into P wave, QRS complex, and T wave. The shape information of those waves is useful for further diagnosis. The interval

parameters (PR interval, QRS interval, QT interval) defined by those five fiducial points are also important for examining a patient's heart conditions. Thus, to evaluate the quality of the reconstructed ECG, we consider both morphological metrics and the accuracy of time interval recovery.

Evaluation of Waveform Morphology: We apply the Pearson correlation (ρ) and relative root mean squared error (rRMSE) as the metrics for evaluating the ECG morphological reconstruction. They are defined as follows:

$$\rho = \frac{(\mathbf{x} - \bar{x})^{\mathrm{T}}(\hat{\mathbf{x}} - \bar{\hat{x}})}{\|\mathbf{x} - \bar{x}\|_2 \left\|\hat{\mathbf{x}} - \bar{\hat{x}}\right\|_2}, \qquad \mathrm{rRMSE} = \frac{\|\mathbf{x} - \hat{\mathbf{x}}\|_2}{\|\mathbf{x}\|_2}. \tag{2.14}$$

where $\mathbf{x}$, $\hat{\mathbf{x}}$, $\bar{x}$, and $\bar{\hat{x}}$ denote the ground-truth ECG cycle, the recovered ECG cycle, and the average of all coordinates of the vectors $\mathbf{x}$ and $\hat{\mathbf{x}}$, respectively.

Evaluation of Time Interval Recovery: Three important ECG interval parameters are studied in this work, including the PR interval, the QRS duration, and the QT 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. We use the segment from P point to R point of ECG as the approximated PR interval. A prolonged PR interval can indicate the possibility of first-degree heart blockage [57]. 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. 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. A prolonged QT interval may lead to ventricular tachycardia [57].

We apply a combination of several established algorithms [104, 118, 119] to detect

Reconstruction Scheme	Configuration	
	Sparsity Constraint?	Linear Mapping Between Representations?
DCT [165]	n.a.	✓
CPDL [85]	n.a.	n.a.
ScSR [156]	ℓ_1	n.a.
SCDL [148]	ℓ_1	✓
CDL [154]	ℓ_0	n.a.
XDJDL (**proposed**)	ℓ_0	✓

Table 2.3: The configuration comparison of the models implemented for ECG reconstruction includes sparsity constraint on the representations and the learnable linear mapping between the representations of PPG and ECG.

the major fiducial points of both the ground-truth ECG and the reconstructed ECG to obtain the above-mentioned interval parameters. We apply the mean absolute error (MAE) in Eq. (2.15) to evaluate the time recovery accuracy:

$$\text{MAE} = \frac{1}{N}\sum_{n=1}^{N}|L_{rec} - L_{ref}|. \tag{2.15}$$

where the L_{rec} and L_{ref} are the interval length (in seconds) of the reconstructed ECG and ground-truth ECG signals, respectively, and N is the total number of cycles for evaluation.

2.4.3 Overall Morphological Reconstruction

We compare our proposed XDJDL method with the state-of-the-art in ECG reconstruction from PPG, which used DCT based method [165]. In addition, we apply several representative and state-of-the-art models of coupled or semi-coupled dictionary learning, including CPDL [85], ScSR [156], SCDL [148], and CDL [154], to compare with the proposed XDJDL method on the ECG reconstruction task. The codes for the prior art are downloaded from the respective authors' websites. The configurations of the prior art

Reconstruction Scheme	ρ			rRMSE		
	$\hat{\mu}$	med	$\hat{\sigma}$	$\hat{\mu}$	med	$\hat{\sigma}$
DCT [165]	0.71	0.83	0.31	0.67	0.60	0.26
CPDL [85]	0.74	0.85	0.31	0.63	0.56	0.35
ScSR [156]	0.82	0.89	0.23	0.54	0.52	0.21
SCDL [148]	0.83	0.89	0.21	0.52	0.49	0.22
CDL [154]	0.85	0.95	0.25	0.49	0.34	0.51
XDJDL (**proposed**)	**0.88**	**0.96**	0.23	**0.39**	**0.29**	0.31

Table 2.4: Quantitative performance comparison for ECG waveform inference.

methods are listed in Table 2.3. The characteristics of these models can be concluded as (1) the way they represent the signals with any sparsity constraints and (2) whether the cross-domain signal representations are assumed to be identical or linearly related by a learnable mapping.

To make a fair comparison, we evaluate the DCT-based reconstruction system in the *subject-independent* training mode where a linear transform $\mathbf{W}_{\mathrm{DCT}}$ is learned using training data from all patients. The normalized PPG/ECG cycle length is chosen as $d = 300$. For XDJDL, the dictionary size for ECG cycles is $k_e = 320$, and the dictionary size for PPG cycles is $k_p = 9000$. The sparsity parameters are set to be $t_e = 10$ and $t_p = 10$. The weights for regularization terms are $\alpha = 1$ and $\beta = 1$. For other dictionary learning models, we have also done the grid-search for hyperparameter tuning to achieve the best performance. We split the data from each patient into training and test sets, and the training data ratio is 80%.

Table 2.4 shows the quantitative comparison of the ECG morphological reconstruction performance. From the statistics of the sample mean, standard deviation, and median of ρ and rRMSE, we can see that our proposed XDJDL method outperforms both the

DCT-based algorithm and other representative coupled/semi-coupled dictionary learning models. Specifically, the average rRMSE is reduced from 0.49 to 0.39, or 20.4% lower than CDL [154], which is the second-best among all competing models.

In Fig. 2.3, we present visualization examples of ECG waveform reconstruction using all the competing models and our proposed XDJDL model. The three patients have different types of disease diagnosis. We observe that even though the waveform variances between the PPGs are relatively smaller than those between the ECGs, our proposed XDJDL method can recover most of the details well in the ECG signal from the PPG signal, suggesting that our method has preserved the intrinsic relation between the atoms from PPG and ECG dictionary pair. In particular, for the second-best CDL [154] method that can reconstruct the overall shape of ECG cycles reasonably well, it has glitches in recovering the details, such as the P wave of the first and last cycle of Patient 2 and the QRS complex of the first cycle of Patient 3.

When the cycle-wise disease information is available, we can apply the proposed label-consistent XDJDL (LC-XDJDL) model from Chapter 2.3.3 to leverage the label information for more accurate monitoring of ECG from the PPG signal. We consider the following scenarios: 1) For cases where the disease information is not directly provided in the test phase, we can first predict that from the PPG signals. Here, we have trained an SVM classifier for the PPG multi-class disease classification and chosen the best hyperparameters with a five-fold cross-validation method. The classification accuracy for the PPG test set reaches 92%. We denote the corresponding label-consistent model as LC1-XDJDL. It will take the predicted labels to build the discriminative representation matrix $\mathbf{Q}$. 2) When we have the ground-truth disease labels in the test phase, we can

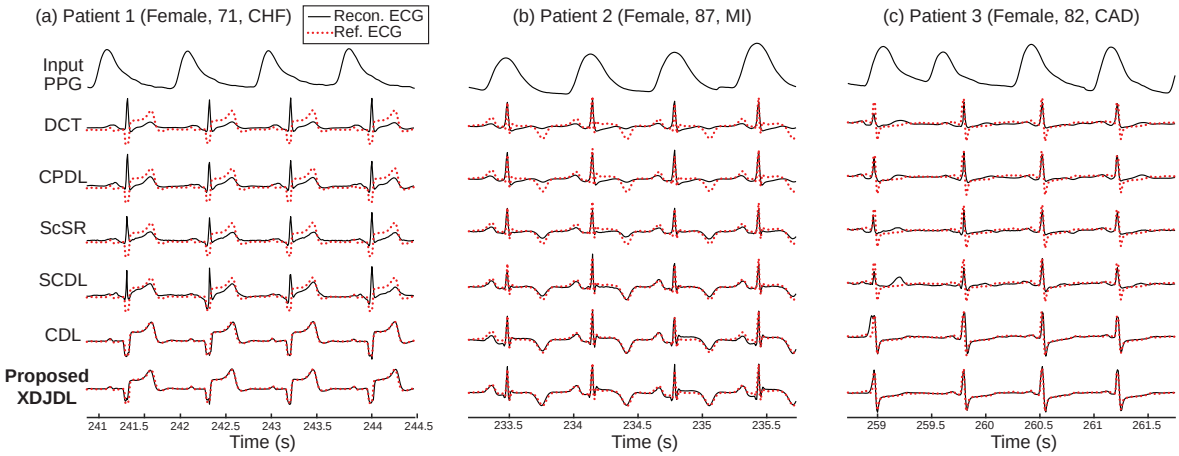


Figure 2.3: Qualitative comparison of the ECG signals inferred by different approaches. Examples are from (a) a 71-year-old female with congestive heart failure, (b) an 87-year-old female with myocardial infarction, and (c) an 82-year-old female with coronary artery disease. From top to bottom: the input PPG signal from which the ECG is inferred in subject-independent mode, results by DCT method [164], CPDL [85], ScSR [156], SCDL [148], CDL [154], and our proposed XDJDL.

leverage that disease information directly as matrix $\mathbf{Q}$ and the corresponding model is named LC2-XDJDL.

We list the comparison of ECG reconstruction performance using the XDJDL, LC1-XDJDL, and LC2-XDJDL models in Table 2.5. On average, the Pearson coefficient improves from 0.88 to 0.90 with the predicted label information, and to 0.92 with the ground-truth disease type as input. The improvement in terms of the rRMSE is also consistent with the Pearson coefficient. In addition to the reconstruction performance improvement, the label-consistent mapping that relates the PPG sparse codes to disease type in LC-XDJDL helps us understand the role of PPG in diagnosis with a rich ECG knowledge base.

Reconstruction Scheme	ρ			rRMSE		
	$\hat{\mu}$	med	$\hat{\sigma}$	$\hat{\mu}$	med	$\hat{\sigma}$
XDJDL	0.88	0.96	0.23	0.39	0.29	0.31
LC1-XDJDL	0.90	0.96	0.20	0.36	0.27	0.28
LC2-XDJDL	**0.92**	**0.97**	0.17	**0.33**	**0.26**	0.25

Table 2.5: Numerical comparison of ECG signal inference performance among XDJDL, LC1-XDJDL, and LC2-XDJDL methods.

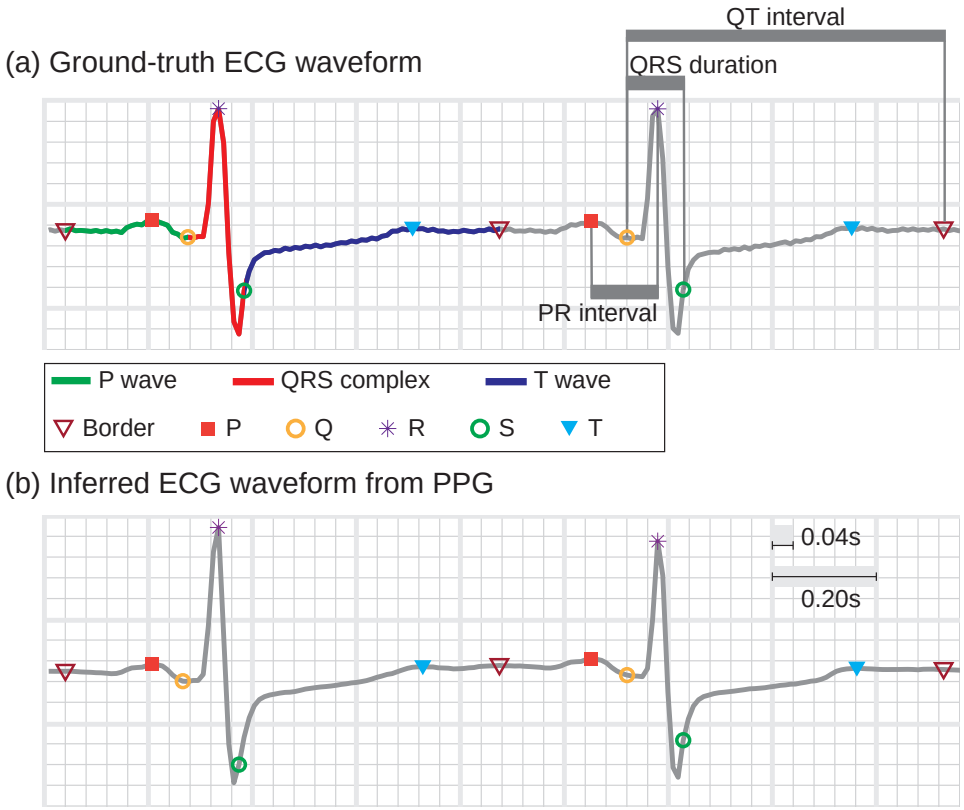


Figure 2.4: (a) shows two cycles of the reference ECG signal and (b) shows two cycles of the inferred ECG signal. In the first cycle of (a), the green curve represents the P wave, the red curve is the QRS complex, and the dark blue curve shows the T wave. The PR interval, QRS duration, and the QT interval are all labeled in the second cycle of (a).

2.4.4 Subwave Morphological Reconstruction

In the above section, we have shown that our proposed XDJDL outperforms the DCT model and other representative dictionary learning models, and its performance can

be better if the disease label (LC-XDJDL) can be utilized for ECG reconstruction and monitoring.

In this section, we zoom into the reconstruction performance of the subwave of ECG cycles using XDJDL and LC-XDJDL methods. Because each subwave refers to different atrial and ventricular depolarization and re-polarization activities, by zooming in, we can have a better idea of how our methods behave on the inference for different phases of the heart activities. A combination of the ECG major point detection algorithms [104, 118, 119] is used to locate P/Q/R/S/T points of ECG waveform, which helps segment the ECG cycle into subwaves for the evaluation of morphological reconstruction.

Fig. 2.4 shows an example of the major points detection results on two cycles of the reference ECG (Fig. 2.4(a)) and the reconstructed ECG (Fig. 2.4(b)) from a patient with coronary artery disease. In this example, we observe that the locations of the detected major points in both signals are close, indicating a good reconstruction of the ECG waveform. We empirically separate the adjacent ECG cycles at a point that splits the neighboring R-R peaks at the ratio of six to four. After that, a complete ECG cycle is divided into three subwaves, including the P wave that starts from the border point on the left of the ECG cycle and ends at the Q point, the QRS complex from the Q to S point, and the T wave from the S point to the right border point. Only a very small portion of reference and reconstructed ECG cycle pairs cannot be detected with a consistent set of fiducial points. The overall number of effective cycles for subwave evaluation is around 92% out of all test cycles, and those effective cycles only have a slightly improved Pearson coefficient (1% on average) compared to the original test dataset.

Table 2.6 lists the reconstruction performance on the three subwaves of the ECG cy-

Reconstruction Scheme	$\bar{\rho}$			$\overline{\text{rRMSE}}$		
	P wave	QRS complex	T wave	P wave	QRS complex	T wave
XDJDL	0.81	0.92	0.84	0.53	0.33	0.41
LC1-XDJDL	0.83	0.93	0.86	0.49	0.30	0.37
LC2-XDJDL	0.86	0.94	0.89	0.45	0.28	0.34

Table 2.6: Comparison of subwave reconstructions in the mean of ρ and rRMSE.

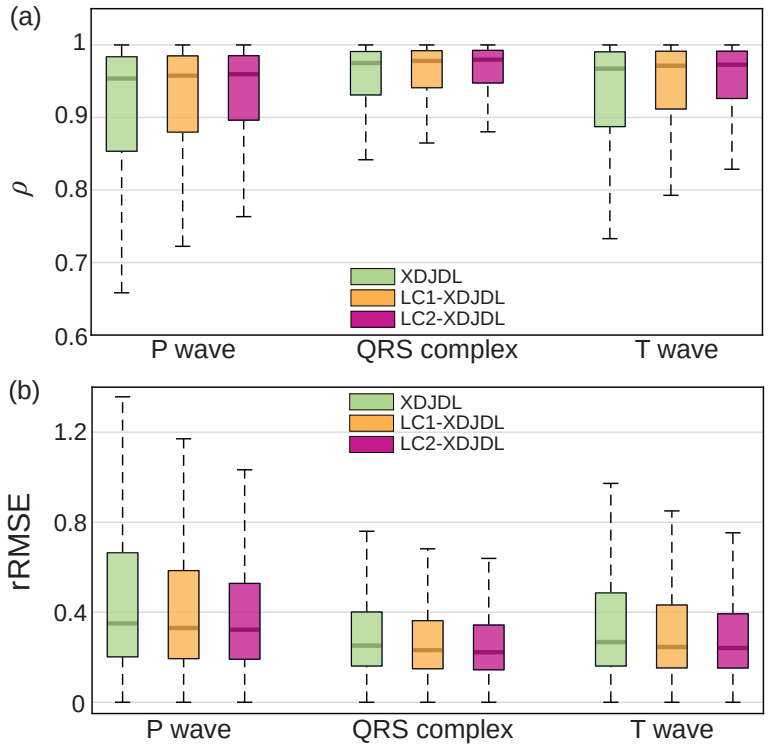


Figure 2.5: Comparison of subwave reconstruction performance across XDJDL, LC1-XDJDL, and LC2-XDJDL models. The statistics of (a) Pearson coefficient ρ and (b) rRMSE are summarized using the boxplots.

cle in terms of the mean of Pearson coefficient and rRMSE using XDJDL, LC1-XDJDL, and LC2-XDJDL models. The comparison of results across models is consistent with the results of the overall comparison in Table 2.5. We also observe that the reconstruction for the QRS complex is better than that for the T wave, which is better than that for the P wave. The mean Pearson coefficient of the QRS complex by LC2-XDJDL is 0.94, higher

than the overall cycle reconstruction of 0.92, while that of the T wave is slightly lower than the overall performance with the mean Pearson coefficient as 0.89 and that of the P wave is 0.86.

In addition to the mean of Pearson coefficient and rRMSE, Fig. 2.5 shows the comparison of the statistics of Pearson coefficient and rRMSE in boxplots for the three subwaves of ECG so that we can see the overall result distribution of the two metrics. We observe that the medians of ρ and rRMSE for each of the three subwaves are very similar across the proposed models. Specifically, the medians of ρ of P wave are 0.95, 0.96, and 0.96, respectively, those of QRS complex are all 0.98, and those of T wave are all 0.97; the medians of rRMSE of P wave are 0.35, 0.33, 0.32, those of QRS complex are 0.25, 0.23, 0.22, and those of the T wave are 0.27, 0.25, and 0.24, respectively. Analysis of these boxplots suggests that our proposed models can preserve the relation between PPG and QRS complex well. The overall reconstruction performance can be improved if the relations between PPG and P and T waves are better learned.

2.4.5 Time Interval Recovery

In addition to the morphological reconstruction evaluation, we evaluate whether the time intervals are well preserved. The labeling of those intervals is shown in Fig. 2.4.

From columns 2-4 in Table 2.7, we can compare the average of the reconstructed intervals and the reference intervals. For PR intervals, the difference between the reconstructed and reference is approximately 4%; for QRS durations, such difference is within 3%; and for QT intervals, the difference is less than 1%. This suggests that, on average,

Reconstruction Scheme	Mean (in seconds)			MAE (in seconds)		
	PR	QRS	QT	PR	QRS	QT
XDJDL	0.164	0.115	0.331	0.030	0.012	0.030
LC1-XDJDL	0.166	0.116	0.331	0.026	0.011	0.027
LC2-XDJDL	0.167	0.115	0.331	0.025	0.010	0.025
Reference	0.172	0.113	0.328	-	-	-

Table 2.7: Comparison of timing interval recovery accuracy in MAE.

the timing information of the intervals is preserved well. From columns 5-7 in Table 2.7, we also notice that the MAEs of the PR interval are 0.030s, 0.026s, and 0.025s using XDJDL, LC1-XDJDL, and LC2-XDJDL models, respectively. The relatively large error in the timing of PR interval recovery is consistent with the result of the P wave reconstruction performance shown in Chapter 2.4.4. Nevertheless, the MAE of the timing for the QRS complex is around 11ms, which is just a quarter of the smallest grid on the conventional hand copy of ECG recorders (40 ms) and is negligible given the sampling rate (125 Hz) of the ECG signal in the MIMIC III dataset. The MAE of the QT interval is around 27ms, which is less than three-quarters of the smallest grid on ECG graph paper and is around 8% of the QT interval (0.331s).

2.5 Discussions

2.5.1 Result Using PPG-based Segmentation Scheme

In Chapter 2.4, we have evaluated our proposed models based on the assumption that the cycle information from ECG signals is available to separate the ECG/PPG time-series signals into training and test cycles. But in practice, we may not have the ground

truth of cycle segmentation from ECG. Thus, we consider such realistic scenarios 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. The PPG onsets are used for segmentation rather than the PPG peaks because of the underlying physiological meaning as we have mentioned in Chapter 2.3.1. For ease of notation, we denote:

- R2R: segmentation scheme based on R peaks of ECG for both training and test data, which is used in Chapter 2.4;
- O2O-1: segmentation scheme based on PPG onsets for both training and test data;
- O2O-2: segmentation scheme based on R peaks of ECG for training data and based on PPG onsets for test data.

Due to the discrepancy between the detected locations of PPG onset and R peak of ECG from the same cycle, the "estimated PPG cycles" using O2O schemes slightly vary from the PPG cycles which are segmented by R2R. To single out the contribution to the ECG reconstruction error due to the discrepancy in the waveform shape rather than the misalignment of the ECG peaks, we evaluate O2O schemes after compensating for the time offset between the reconstructed ECG and original ECG signals. This is done by shifting each reconstructed ECG cycle in time so that the original and reconstructed ECG signals are matched according to their R peaks. The comparison result is shown in Table 2.8. Compared to R2R, when using the O2O-1 scheme, the average Pearson coefficient drops from 0.88 to 0.70, and the average rRMSE rises from 0.39 to 0.66. And using the O2O-2 scheme can help improve the performance compared to O2O-1, where the mean Pearson coefficient becomes 0.80 and the mean rRMSE becomes 0.55.

Reconstruction Scheme	ρ			rRMSE		
	$\hat{\mu}$	med	$\hat{\sigma}$	$\hat{\mu}$	med	$\hat{\sigma}$
XDJDL (O2O-1)	0.70	0.84	0.32	0.66	0.57	0.39
XDJDL (O2O-2)	0.80	0.88	0.24	0.55	0.48	0.32
XDJDL (R2R)	0.88	0.96	0.23	0.39	0.29	0.31

Table 2.8: Quantitative Comparison of different segmentation schemes.

2.5.2 Evaluation on the Capnobase TBME-RR Dataset

In this section, we experimented with the Capnobase TBME-RR database [74] that contains forty-two eight-minute sessions from 29 children and 13 adults during elective surgery and routine anesthesia. Each session corresponds to a unique participant and contains simultaneously recorded PPG and ECG signals. The signals are recorded with a sampling frequency of 300 Hz. The dataset covers a wide range of participants' ages, which is from one-year-old to sixty-three-year-old with the median age being fourteen. Thus, this dataset is used for a supplementary evaluation of the proposed method from the angle of age variety in addition to disease variety in the mini-MIMIC-33 dataset.

We first pruned the signals according to the artifact labels provided in the dataset and preprocessed the signals using the method in Chapter 2.3.1 to obtain aligned and normalized signal pairs. To be consistent in the evaluation, like what we did in Chapter 2.4, we selected the first 80% of the data from each subject as the training set and the rest for testing.

Table 2.9 summarizes the performance comparison using the Capnobase TBME-RR dataset. Our proposed XDJDL method outperforms all the other groups in terms of the mean and median rRMSE by a large margin. Even though the CDL [154] method is

Reconstruction Scheme	ρ			rRMSE		
	$\hat{\mu}$	med	$\hat{\sigma}$	$\hat{\mu}$	med	$\hat{\sigma}$
DCT [165]	0.902	0.919	0.066	0.427	0.413	0.128
CPDL [85]	0.956	0.968	0.049	0.282	0.247	0.150
ScSR [156]	0.967	0.976	0.039	0.286	0.247	0.165
SCDL [148]	0.971	0.978	0.038	0.191	0.166	0.101
CDL [154]	**0.980**	**0.991**	0.062	0.219	0.145	0.296
XDJDL (**proposed**)	0.979	0.990	0.048	**0.146**	**0.105**	0.122

Table 2.9: Quantitative performance comparison for ECG waveform inference using the Capnobase TBME-RR database.

0.1% better than our proposed method in mean and median correlation coefficient ρ, our method achieves a 26% smaller $\hat{\sigma}$ of ρ than the CDL method, showing that our proposed method achieves a good performance of ECG reconstruction more consistently for all participants.

2.5.3 Feasibility Analysis of The Proposed Method for The Internet-of-Healthcare-Things (IoHT)

In this section, we analyze two important practicality issues when applying our proposed ECG reconstruction techniques to healthcare IoT devices. One issue is energy consumption. The sensors used to capture physiological signals, e.g., PPG signals, are mostly wearable devices, which are powered by batteries [55]. Thus, being energy-efficient is necessary to ensure continuous signal acquisition, data transmission, and monitoring. The other issue is computational cost. As is mentioned in [9, 55], applications that require lower latency need higher computational capabilities. Thus, the computational load of the algorithms needs to be considered in real-world scenarios.

The first issue about energy consumption in wearable devices can be resolved by

the existing mature technologies like the Bluetooth low-energy module commonly applied for low-power wireless communication in wearable healthcare devices [153]. In the test phase of our proposed XDJDL and LC-XDJDL frameworks, PPG signals acquired by the wearable devices can be transmitted to the IoT devices, such as smartphones, at low power with the help of those modules. For the second issue about computational cost, with the dictionary pairs constructed locally and stored in the cloud or edge devices, the computational cost is mainly from sparse coding and lightweight matrix multiplication. Since sparse coding via OMP in our proposed methods is proven to be able to be executed on the IoT platform in real-time [6], we envision that our proposed frameworks can satisfy the practical requirements well.

To further evaluate quantitatively the feasibility of applying our proposed method to IoHT platforms, we examine the following metrics to measure the usage of computational resources to reconstruct one ECG cycle:

1. Computational time
2. Memory space
3. Energy consumption

The specifications of the laptop we used for the experiment are as follows: Processor: i7-8650U; Architecture: Intel x86; CPU Frequency: 1.90GHz; Cores: 4; RAM: 24GB. Our test here is designed to resemble an online inference scenario in which new sequences of continuous ECG waveform need to be inferred by the IoHT system with the input PPG waveform. The experiment is repeated 100 times to evaluate the memory space and the

Reconstruction Scheme	Computational Time (ms)	Memory Space (MB)	FLOP Consumption
XDJDL (proposed)	15.7 ± 0.9	31.4 ± 1.3	60.2M

Table 2.10: Computational resources consumed to reconstruct test ECG cycles using the proposed XDJDL method.

average computational time for each cycle. Note that the actual energy consumption estimation can be complex, as it depends on the operating system, the temperature inside and outside the device, and the efficiency of the power supply. Thus, we use FLOP (Floating-point Operations) here as the measure for energy consumption, as it is independent of hardware configurations given the algorithm. With FLOP, the energy in joule can be estimated as it is proportional to FLOP given FLOPS (FLOP per Second) per watt, i.e., FLOPS/W, specified by the IoT device.

We list the computational resources consumed by the proposed XDJDL method in Table 2.10. The average computational time to reconstruct each ECG cycle is 15.7ms, which is one to two orders of magnitude shorter than a heart cycle (around 0.5s to 1s per beat at rest), suggesting that the processing can be done in real-time. In addition, the 31.4 MB memory space and 60.2 MFLOP required by the proposed XDJDL method are well within the capability of such commonly seen IoT platforms as the Raspberry Pi 3B (RAM: 1 GB, 0.73 GFLOPS/W) [5] for the research prototype that has not been optimized for deployment. Considerable reductions in computing resources are possible with industry-grade implementation.

2.5.4 Limitations of The Proposed Method

2.5.4.1 Performance of Leave-One-Out Experiment

As a proof of concept and considering the current moderate amount of available data, we have so far split each patient's data into training and test sets. This corresponds to the trend of "precision medicine" to tailor the healthcare practice to individual patients. Meanwhile, we are curious how the algorithm would behave if the test patient is never seen in the training phase, corresponding to the situation of training models for the whole population or patient groups categorized by gender, age, race, or other ways. We will examine this through leave-one-out experiments.

We apply a pre-clustering process based on the ECG data to select a sub-group of patients with similar ECG features for the leave-one-out experiment. First, we reduce the dimension of the ECG cycles by principal component analysis (PCA), and then we use K-means to cluster the ECG features after PCA. Based on the clustered ECG features, we select the largest cluster of ECGs from 19 patients. The mean Pearson coefficient for the leave-one-out experiment on the 19 patients is 0.74 (std: 0.15, median: 0.77).

From the result, we can see that as expected, the leave-one-out experiment is a more challenging case given the large variability of ECG data morphologies of ICU patients and the limited number of patients in the collected dataset. Based on the results in Chapter 2.4.3, we see the encouraging capability of recovering large variations in ECG from relatively small variations in PPG across cycles and patient populations. This suggests a strong potential for predicting ECG from PPG of unseen patients through further research

and larger data collection. In our follow-up work, we are considering an improved problem definition and data collection procedure to enhance the generalization capability of learning.

2.5.4.2 Performance Evaluation on A Motion Dataset

So far, we have demonstrated the feasibility and improved accuracy of ECG waveform inference from PPG using the proposed methods on two benchmark datasets [71,74] in Chapter 2.4.3 and Chapter 2.5.2. Those datasets were collected under a resting condition with relatively small movement artifacts. Noises and artifacts were still present in those datasets but in a controlled manner, which leads to good quality of data acquisition and is beneficial for the feasibility study and accuracy improvement of reconstructing ECG from PPG. In this section, we consider a more challenging scenario where IoHT devices are worn during exercise and show the preliminary results with the motion-contaminated signals.

Dataset Description:

We adopt the 2015 IEEE Signal Processing Cup dataset [161] for evaluation, which consists of paired PPG and ECG signals from 13 participants during physical exercises. This dataset provided by the Samsung Research Lab in the U.S. aimed to facilitate the study of accurate heart rate (HR) monitoring of PPG signals from wrist-type sensors and included ECG signals as a reference. The PPG signals were collected from the wrist while the subjects ran on a treadmill at speeds of 6 km/h, 8 km/h, 12 km/h, or 15 km/h, respectively. Simultaneously, the ECG signals were collected from the chest and the

acceleration signal was recorded from the wrist by a three-axis accelerometer. All signals were sampled at 125 Hz. Each subject ran once and the total length of the recording was 5 minutes per subject.

Dataset Preprocessing With HR Reference and Adaptive Filtering:

Since the quality of PPG signals is crucial to ECG reconstruction, we first use the absolute error of the PPG estimated HR as a metric to exclude the participants with extremely corrupted PPG signals. Because HR represents the frequency characteristic of PPG that affects the accuracy of determining a PPG cycle. The HR is estimated from the PPG by a state-of-the-art adaptive multi-trace carving (AMTC) [163] algorithm that tracks the HR from the spectrogram of PPG by dynamic programming and adaptive trace compensation. The reference HR values are given in the dataset. Three out of the thirteen participants are excluded as their HR estimation error is quite off likely due to data collection issues and the remaining ten participants' data are used for learning and testing the XDJDL model.

In addition, to improve the quality of noise-contaminated PPG, we conducted recursive least square (RLS) adaptive filtering [75]. We view the contaminated PPG as the sum of the underlying cleaned PPG and motion-induced noise. Suppose the motion artifact corrupted PPG signal at time $n \in [1, N]$ is $\mathbf{p}(n) = \mathbf{d}(n) + \mathbf{m}(n)$, where $\mathbf{d}(n)$ is the underlying cleaned PPG and $\mathbf{m}(n)$ is the noise introduced by motion that is unknown and can be modeled and estimated with the acquired accelerometry signals $\mathbf{a} = [\mathbf{a_x}; \mathbf{a_y}; \mathbf{a_z}]$. In this way, the estimated $\hat{\mathbf{m}}$ may be subtracted from $\mathbf{p}$ to form an estimate of $\mathbf{d}$ for PPG motion artifact compensation. The block diagram for the RLS adaptive filter structure is shown in Fig. 2.6 and the RLS algorithm is described in Algorithm 2 where the object

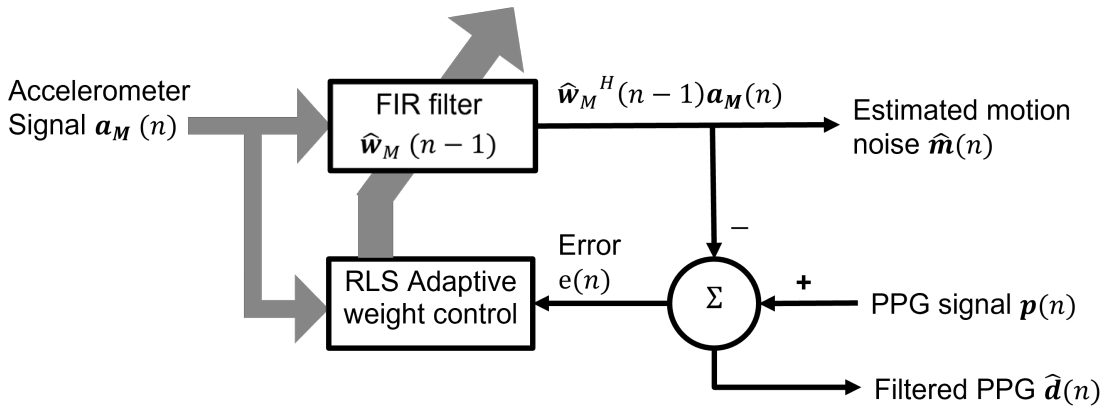


Figure 2.6: Block diagram for RLS algorithm.

function is $\xi = \sum_{n=0}^{N} \lambda^{N-n}|e(n)|^2$, in which $e(n)$ is the prior estimation error of $\mathbf{p}$ by the RLS adaptive filter and λ is the forgetting factor that is set to be 1, i.e., assuming infinite memory. Since there are three channels of accelerometer data, we adopt a *series processing* method in which the raw PPG is denoised with $\mathbf{a_x}$ first, then $\mathbf{a_y}$ is used to denoise the PPG after $\mathbf{a_x}$, and lastly, $\mathbf{a_z}$ is input into the RLS to further denoise the PPG signal after $\mathbf{a_x}$ and $\mathbf{a_y}$.

Algorithm 2 RLS algorithm [60]

Variables: $\hat{\mathbf{w}}_\mathbf{M}$ is the M-tap weight vector; $\mathbf{P}$ is the *inverse of the correlation matrix*; $\mathbf{k}$ is the *gain vector*; $\mathbf{a_M}$ is the input accelerometer data in an M-length window; $\mathbf{p}$ is the raw PPG signal to be filtered.

Initialize: $\hat{\mathbf{w}}_\mathbf{M}(0) = \mathbf{0}$; $\mathbf{P}(0) = \delta^{-1}\mathbf{I}$

for $n = 1, 2, \ldots$ **do**

$\mathbf{k}(n) = \frac{\mathbf{P}(n-1)\mathbf{a_M}(n)}{\lambda + \mathbf{a_M}^{\mathrm{H}}(n)\mathbf{P}(n-1)\mathbf{a_M}(n)}$

$e(n) = \mathbf{p}(n) - \hat{\mathbf{w}}_\mathbf{M}{}^{\mathrm{H}}(n-1)\mathbf{a_M}(n)$

$\hat{\mathbf{w}}_\mathbf{M}(n) = \hat{\mathbf{w}}_\mathbf{M}(n-1) + \mathbf{k}(n)e^*(n)$

$\mathbf{P}(n) = \lambda^{-1}\mathbf{P}(n-1) - \lambda^{-1}\mathbf{k}(n)\mathbf{a_M}{}^{\mathrm{H}}(n)\mathbf{P}(n-1)$

end for

return $\hat{\mathbf{w}}_\mathbf{M}, e$

In the next part, we will compare the ECG reconstruction performance from PPG signals before and after denoising.

Experimental Performance:

Fig. 2.7(a) shows the comparison of the statistics of Pearson coefficient and rRMSE in boxplots for ECG reconstructed from the PPG signal without denoising (referred to as "raw PPG") and RLS filtered PPG signal (referred to as "cleaned PPG"). The average Pearson coefficient of the reconstructed ECG using raw PPG is 0.49 (median: 0.69, std: 0.51) and using cleaned PPG is improved to 0.61 (median: 0.72, std: 0.37). This improvement can be attributed to that the spurious peaks and waves in the motion-contaminated PPG are removed by the RLS filtering. While the noise due to motion is mitigated, distortions in PPG and even ECG waveforms are still present as shown in Fig. 2.7(b). Treating potentially corrupted ECG as the reference and distorted PPG as the input might misguide the learning system, and produce unreliable waveform reconstruction.

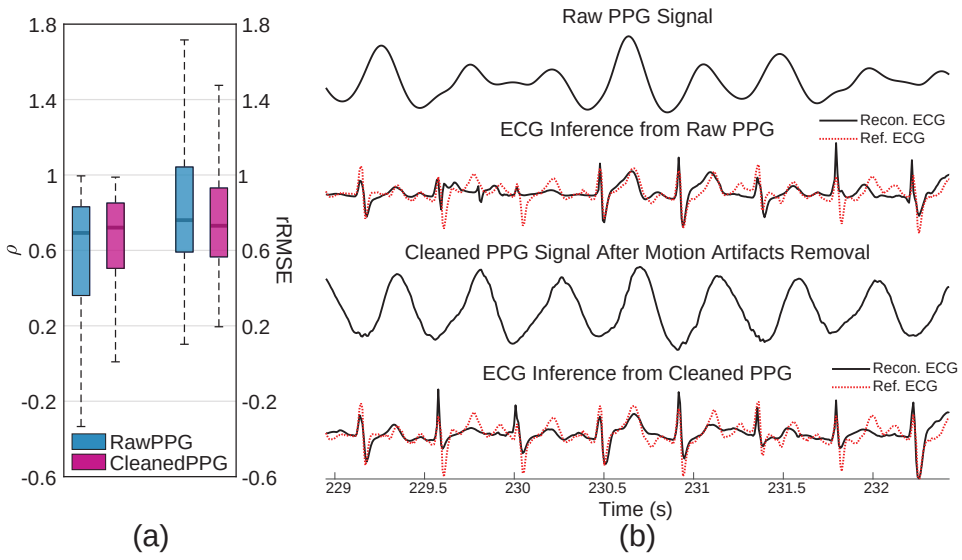


Figure 2.7: (a) Statistical distribution of Pearson coefficient (ρ) and rRMSE for reconstructed ECG from PPG signals before denoising ("raw PPG") and PPG signals after denoising ("cleaned PPG"). (b) Qualitative comparison of raw PPG, cleaned PPG, and the ECG signals inferred from them.

Fig. 2.7(b) shows an example with close to average performance. We observe that

on one hand, the cleaned PPG has clearer cycle shapes than the raw PPG; and on the other hand, some of the physiological characteristics representing the blood flow process are irregular after RLS motion artifacts removal, such as the peak in the third cycle and the ascending and descending slopes in the fifth cycle. Also, the reference ECG signals contain varying ST segment elevations over consecutive cycles during motion. We expect such limitations can be addressed with the development of more advanced PPG and ECG denoising and waveform preserving approaches for preprocessing and the availability of a larger dataset under different types of activities (such as walking, running, driving, climbing stairs, etc).

2.5.5 Future Work Towards Explainable AI

Our proposed XDJDL and LC-XDJDL models accomplished to infer the ECG based on PPG by leveraging the biomedical and statistical relationship between the signals. This is an initial effort to demonstrate a potential benefit from our "explainable" AI, rather than black-box data-driven AI, to provide more user-friendly PPG measurements inferred ECG data for the medical professionals to interpret and offer medical insights. Our framework also helps transfer the rich ECG knowledge base from decades of medical practice to augment the PPG diagnosis for public health.

Given the challenge of making the ECG inference more accurate for an unseen group of subjects, e.g., by age, gender, or other medical and health condition, we are extending our current work with a neural network to further enrich the representation and learn the relation when sufficient data is available. Our ongoing efforts have been

focused on both developing a data collection pipeline for more diversity and coverage of training data and exploring an explainable generative model with strong expressive power to improve the generalization performance. With the step-by-step capturing of the complex models by utilizing the biomedical, statistical, and physical meanings, as well as harnessing the power of the data, we aim to provide explainable AI with our ongoing efforts.

2.6 Chapter Summary

We have proposed a cross-domain joint dictionary learning (XDJDL) framework and the extended label-consistent XDJDL (LC-XDJDL) model for ECG waveform inference from the PPG signal. Compared to the prior art using the DCT method, our proposed method better leverages the data to improve data representation while extending over a model-based approach. The promising experimental results validate that our proposed models can learn the relation between PPG and ECG and reconstruct ECG well. From the analysis for subwave reconstruction and timing of interval recovery, we observe that we can restore the QRS complex and the QT interval in high precision, which is essential for ECG monitoring and to gain more PPG-based diagnosis knowledge. This work reveals the potential of long-term and user-friendly ECG screening from the PPG signals that we can acquire from the daily use of low-cost, low-power wearable devices for IoT and digital twins applications in healthcare.

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 [115] and clinical data suggest that the susceptibility to outcomes of COVID-19 is strongly associated with CVD [99]. 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) [46] 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 [165]). 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.

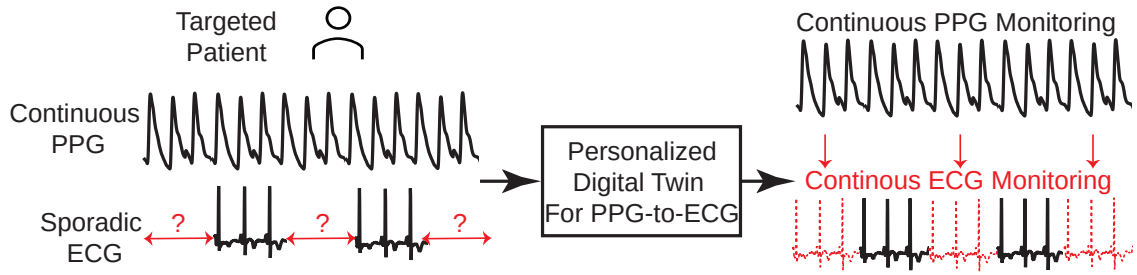


Figure 3.1: Our goal in this work is to build a personalized digital twin model for a targeted patient with his/her limited sporadic paired PPG and ECG cycles, such that his/her ECG can be faithfully and continuously inferred from the continuous PPG measured by the daily wearable devices.

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 [74]	MIMIC-III [71]	BIDMC [107]	SPCup 2015 [161]
Zhu et al. [165]	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. [137]	Same as Zhu et al. [165]	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. [88]	n/a	Same as Tian et al. [137]	8-min from each of the 53 patients; 80%:20% train/test split from each patient	Same as Tian et al. [137]
Vo et al. [147]	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. [28]	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 [88, 136, 137, 147, 164, 165], including the work presented in Chapter 2. Among those works, Vo et al. [147] used randomly selected 8-minute-long PPG-ECG signals from 276 patients in the MIMIC-II database [49] to analyze their models. Zhu et al., 2019 [164]; Zhu et al., 2021 [165]; Tian et al., 2020 [136]; Tian et al., 2021 [137], and Li et al. [88] 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 [165] 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 [88, 136, 137, 147, 164, 165] dedicated to the PPG-based ECG inference problem summarized in Chapter 3.2, the pilot study [164] 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 [164, 165] 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 [88, 147]. 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 convinced 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, 89]. 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.

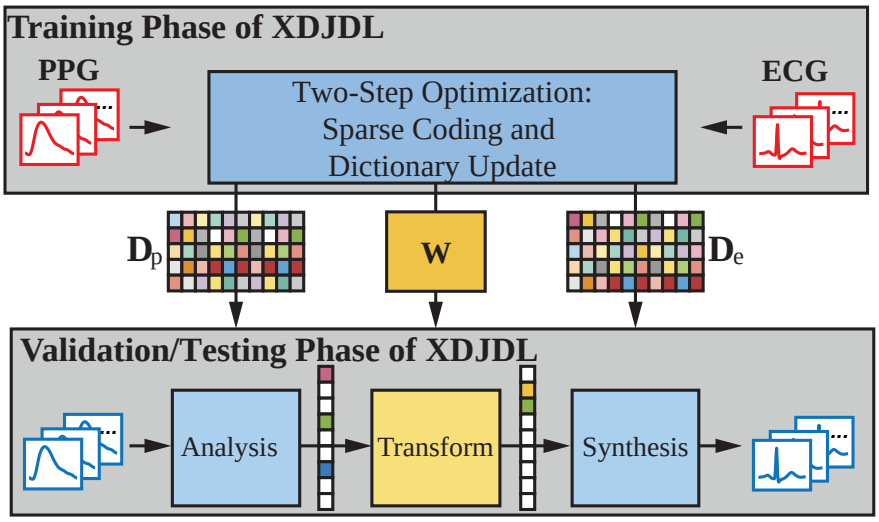


Figure 3.2: The XDJDL model proposed in Chapter 2 is adopted here as the backbone model for ECG inference from PPG. The dictionary pair $\mathbf{D}_p$, $\mathbf{D}_e$, and the linear mapping $\mathbf{W}$ are learned during the training phase, which are later applied to infer ECG from PPG in the validation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

3.3.2 Transfer Learning for Building Precision Healthcare Digital Twins

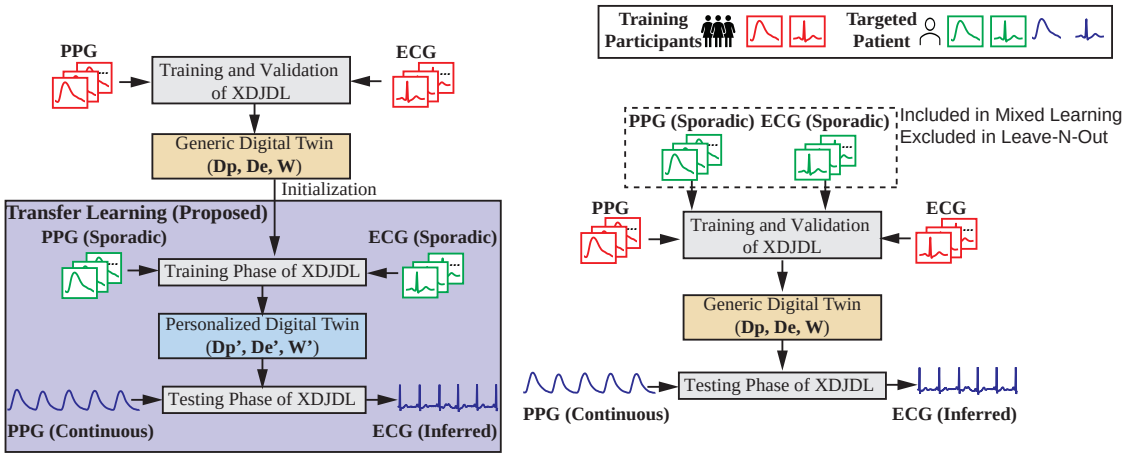
Considering a group of people with abundant paired PPG and ECG signals from their in-hospital stay or annual physical examinations, we denote their corresponding PPG and ECG datasets as $\mathbf{X}_p \in \mathbb{R}^{d\times N}$ and $\mathbf{X}_e \in \mathbb{R}^{d\times N}$, respectively. Each column 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 [132], AliveCor [73], Zio patch [36], and Empetica E4 watch [40]. 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-



(a) Transfer Learning (Proposed) (b) Baseline Comparisons

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

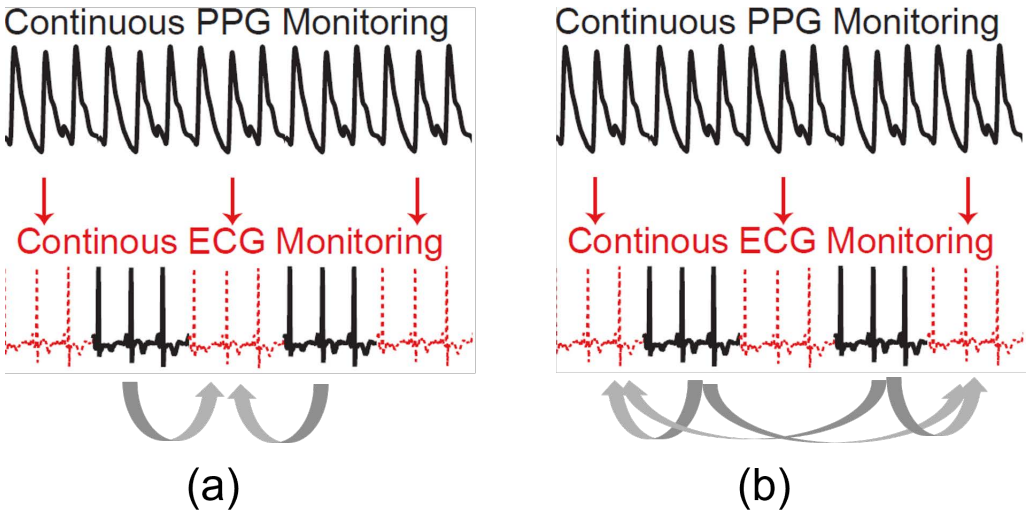


Figure 3.4: The two testing modes that we examine for the learned digital twin model. (a) Interpolation mode where we can "rewind" the ECG during its detachment between two sporadic time stamps that contain known paired PPG and ECG signals. (b) Extrapolation mode where we can check the past ECG or predict 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) [71] 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.

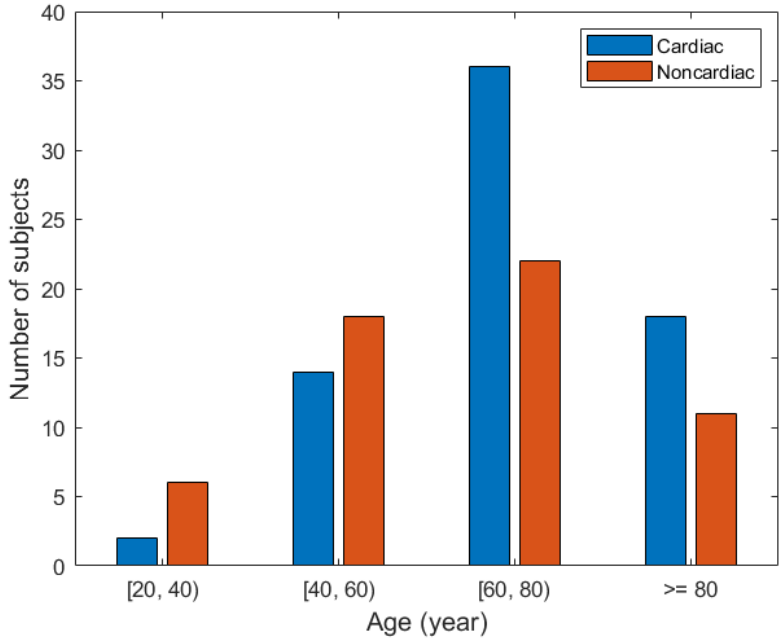


Figure 3.5: Distribution of the 127 patients collected from the MIMIC-III database in different age groups and disease types (mini-MIMIC-127 dataset). Within each age group, the patients with cardiovascular-related diseases are marked in blue on the left, and the patients with non-cardiovascular-related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

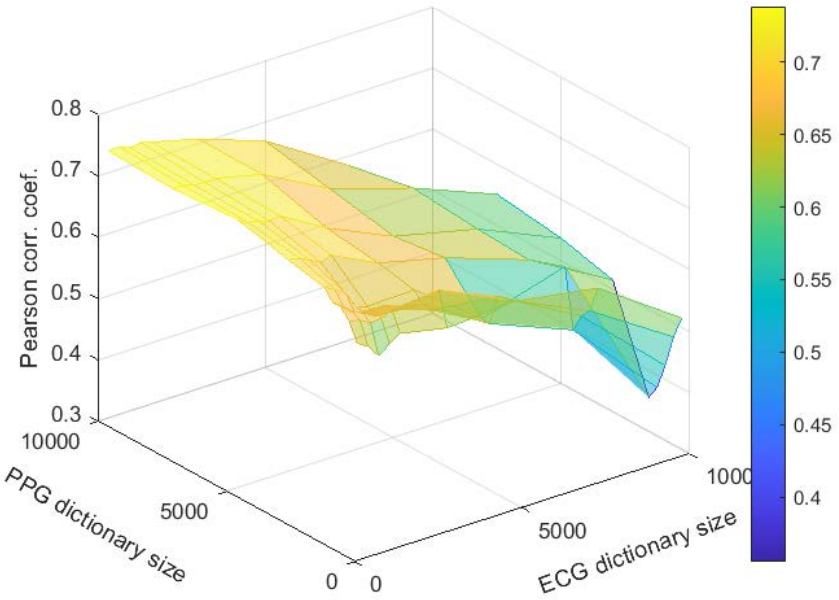


Figure 3.6: The validation performance in terms of Pearson correlation coefficient with respect to different combinations of PPG and ECG dictionary sizes.

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 counter-intuitive 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 [27, 111, 112].

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.

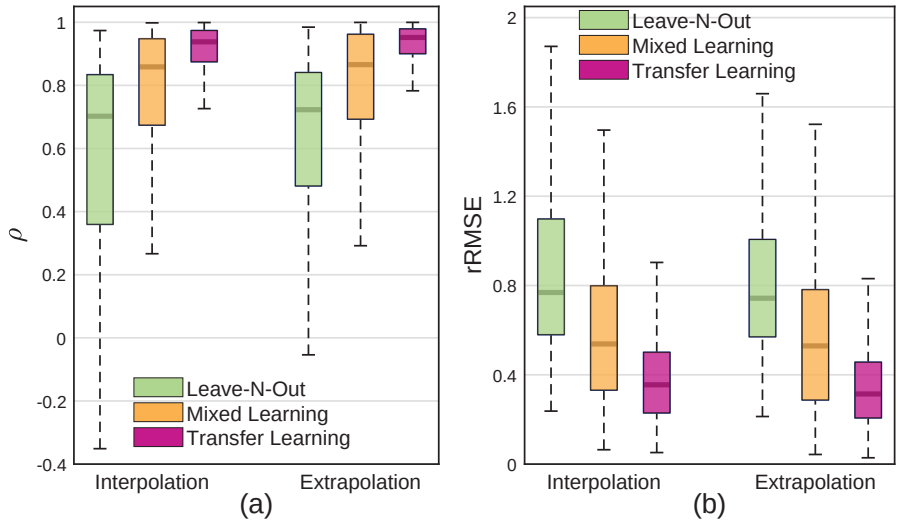


Figure 3.7: Statistical distribution of (a) Pearson correlation coefficient (ρ) and (b) rRMSE for the inferred ECG signals in both interpolation and extrapolation testing modes using different training modes (leave-N-out, mixed learning, and transfer learning).

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

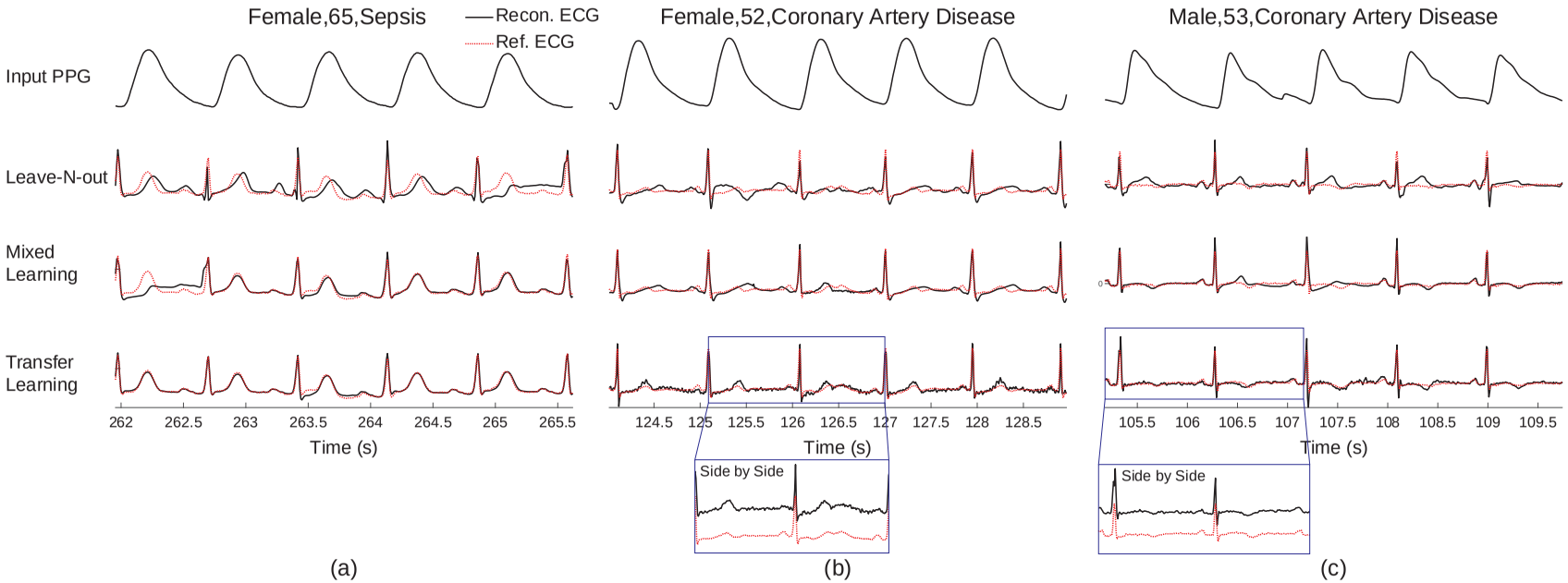


Figure 3.8: Qualitative comparison of the ECG signals inferred in different modes. Examples are from (a) a 65-year-old female with sepsis, (b) a 52-year-old female with coronary artery disease, and (c) a 53-year-old male with coronary artery disease. From top to bottom: the input PPG signal from which 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.

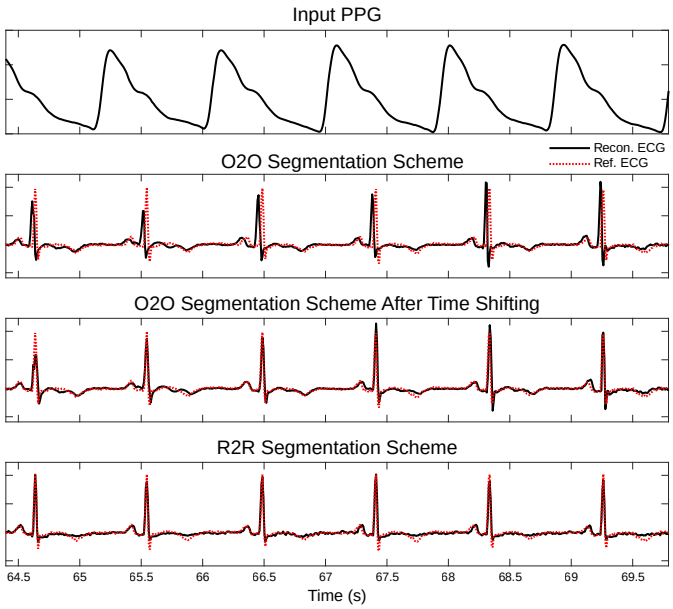


Figure 3.9: Qualitative comparison for different segmentation schemes. From top to bottom: the input PPG signal from which ECG is inferred, results from the O2O segmentation scheme, results after shifting the O2O inferred cycle in time to align the R peaks of the inferred ECG and the reference ECG, and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

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.

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:

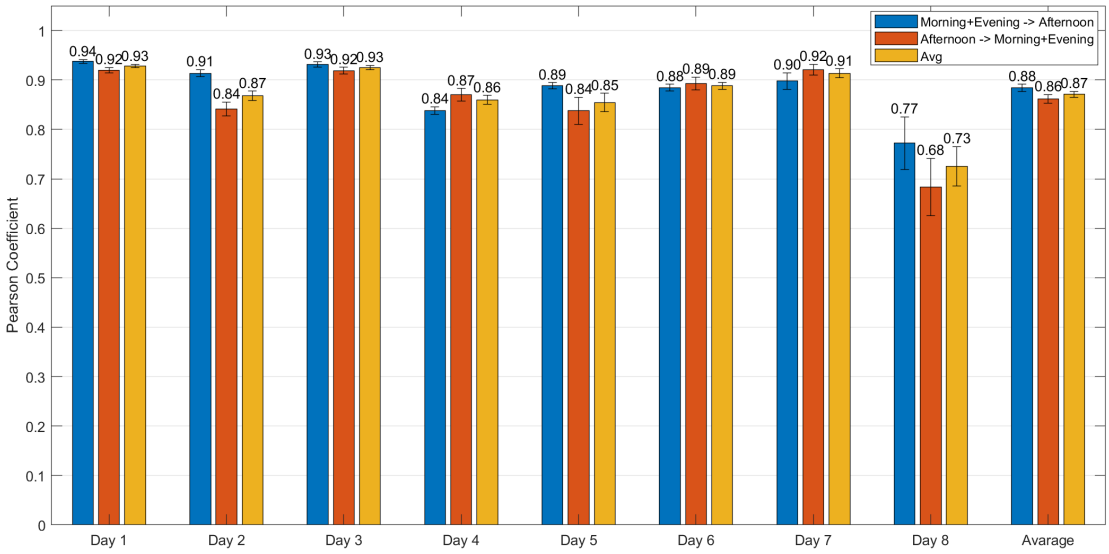
Table 3.5 summarizes the overall performance from each of the learning and evaluation schemes, excluding the training and validation sessions from Day 1 morning, Day 3 morning, Day 6 evening, and Day 8 evening for a fair comparison. We observe that the personalized digital twin updated by the Day 1 morning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

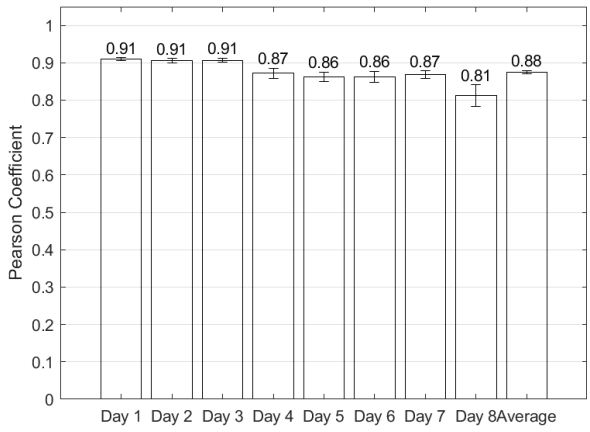
	Scheme (a)	Scheme (b)	Scheme (c)
ρ	0.87 (0.20)	0.88 (0.16)	0.88 (0.22)
rRMSE	0.49 (0.32)	0.49 (0.23)	0.44 (0.26)

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.

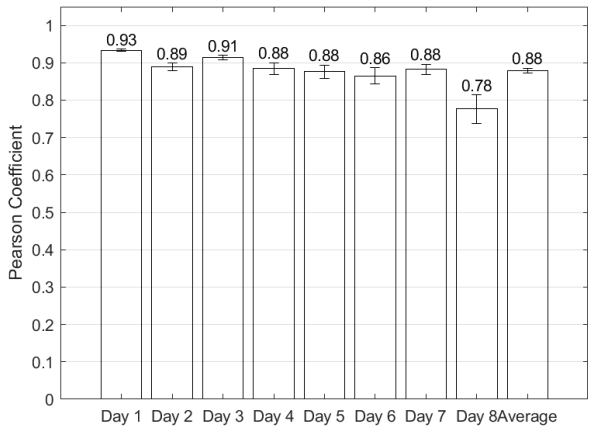
The breakdown of everyday performance in terms of the Pearson coefficient from the three schemes is shown in Fig. 3.11. The height of each bin shows the average correlation coefficient ρ of ECG reconstruction results from the overlapped test sessions of the three schemes each day. Each error bar corresponds 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 better performance than the other two schemes, suggesting that the inference within a



(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.

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) [76]. 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 the Neural network 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 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

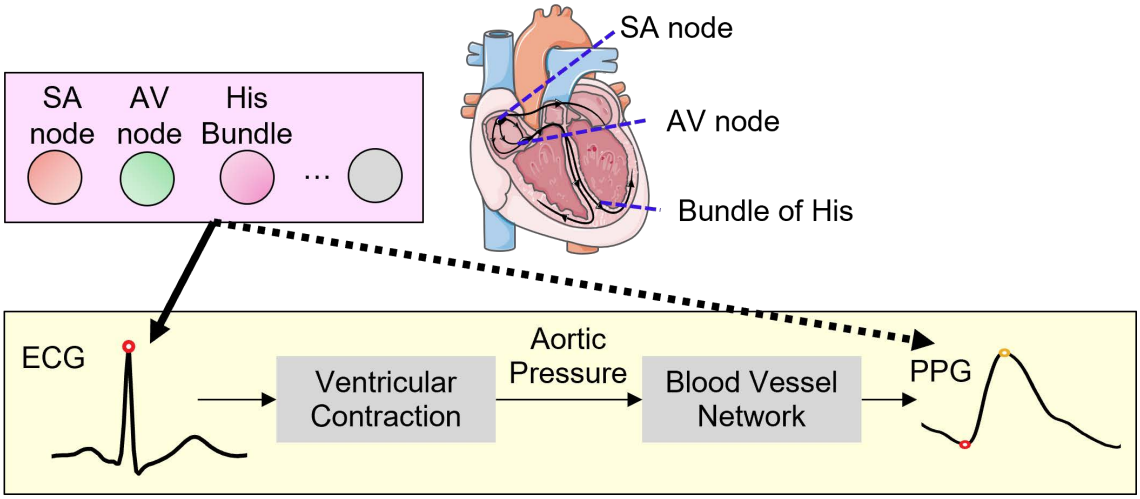


Figure 3.12: The ECG and PPG signal generation paths during heartbeats considering the originating impulses from the heart.

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 [165], XDJDL-based [137] (Chapter 2), and autoencoder-based frameworks [88], 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.

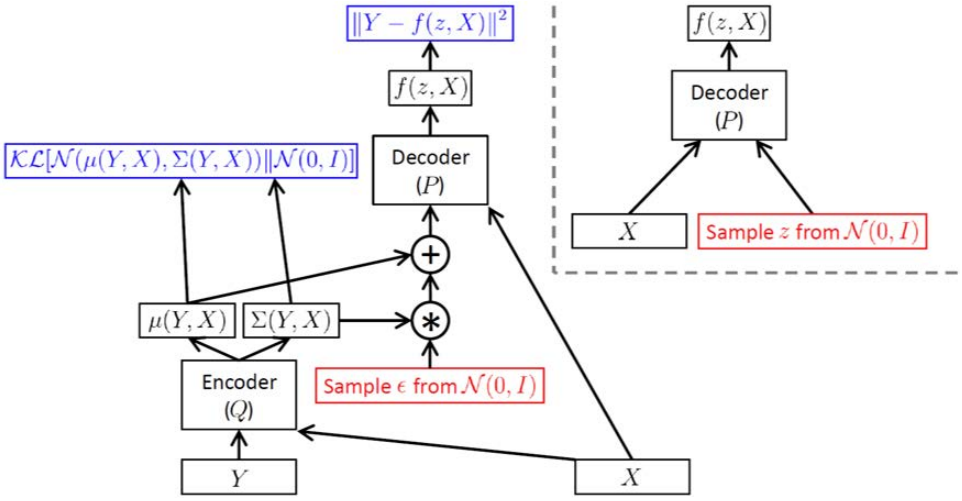


Figure 3.13: The left panel shows the CVAE structure implemented as a feed-forward neural network during the training process. The upper right panel shows the model at test time when we want to sample from $P(Y|X)$. The illustration is adopted from [38].

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.

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

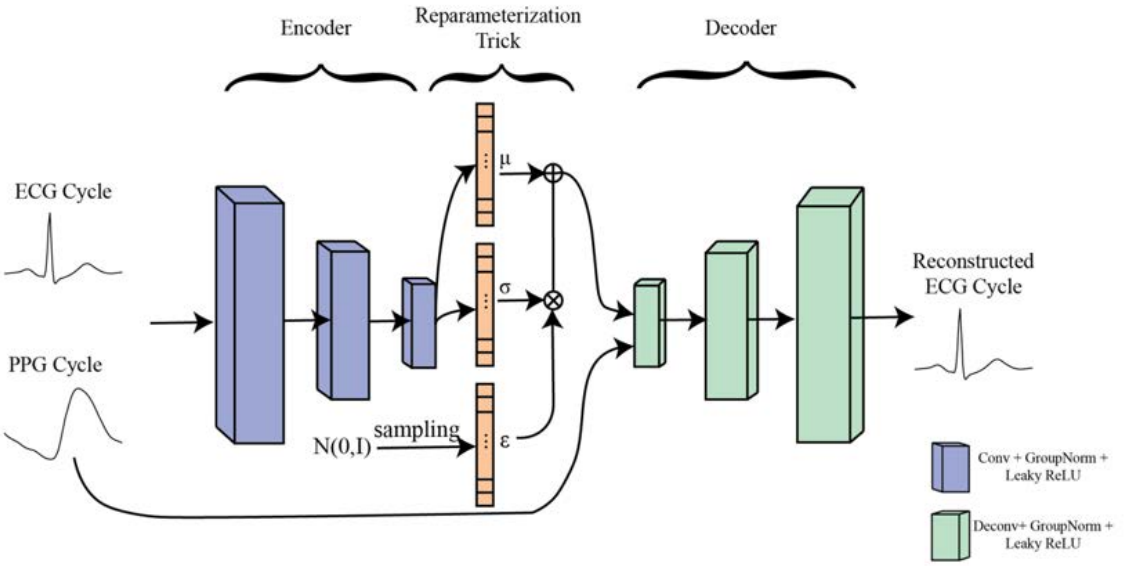


Figure 3.14: The vanilla CVAE model as the backbone for ECG inference from PPG.

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.

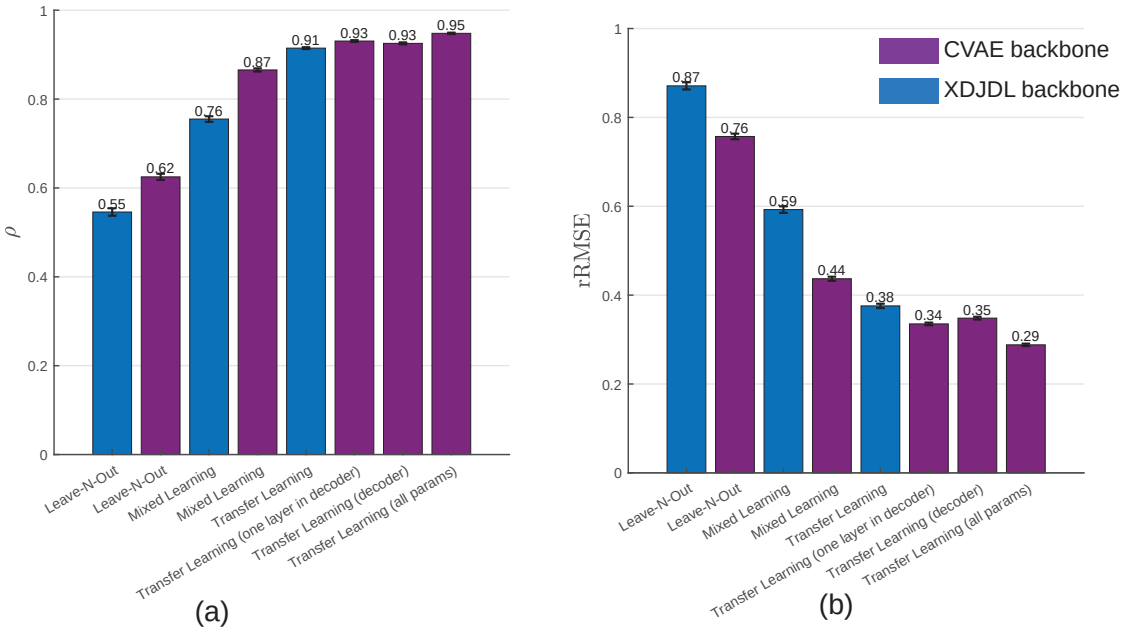


Figure 3.15: The overall performance comparison using XDJDL and CVAE as the backbone models for PPG-to-ECG inference in different training modes, including leave-N-out, mixed learning, and transfer learning. The error bars correspond to the 95% confidence intervals.

Fig. 3.15 presents the comparison of the overall results between using XDJDL and

	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.

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% 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)

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 build-

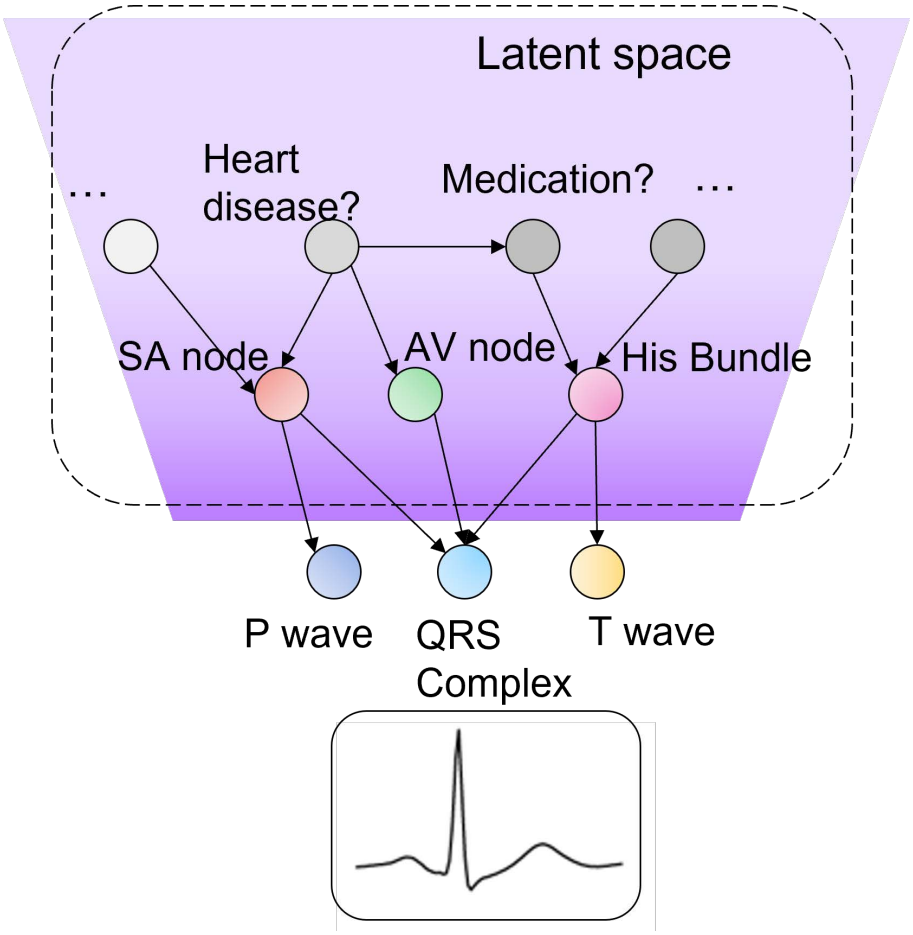


Figure 3.16: Illustration of causal representation learning for ECG inference. The factors that form a causal mechanism in the latent space are assumed to generate the higher dimensional data of the ECG waveform in the observed space. The arrow indicates that the parent node causes the child one.

ing 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 [42, 59] and physiological signal sensing [15, 29]. 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 [151]. 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 counter-intuitive 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 counter-intuitive results and make the models more interpretable and it is drawing increasing attention nowadays to take the advantage of both fields [116].

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

Consider a set of observable 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. A functional causal model is proposed in [106] 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 SEM is a linear 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 $\mathbf{A} = [\mathbf{A}_1|...|\mathbf{A}_n] \in \mathbb{R}^{n \times n}$, where A_{ji} represents the casual strength from Node j to Node i ($A_{ji} = 0$ if there is no causal edge from Node j to Node i). Then the 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 [106], 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 SEM and (2) setting $X_j = x_j$ in the right-hand sides of the other equations of a causal structure in SEM. By investigating the mapping from x to $P(y|do(x))$ for all x, the causal effect of X on Y can be examined.

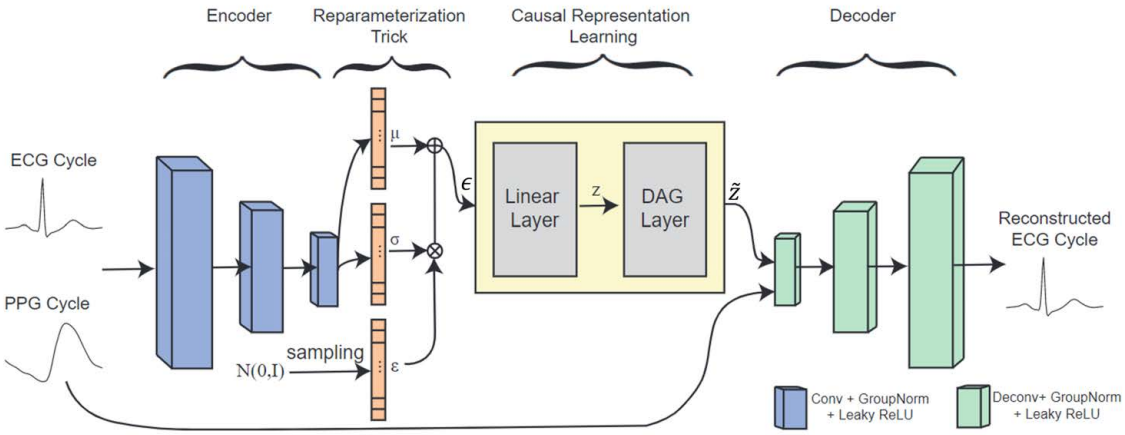


Figure 3.17: The proposed causal CVAE architecture. Compared to the vanilla CVAE structure in Fig. 3.14, the causal CVAE model incorporates the causal representation learning module that helps to learn the causal latent vector $\mathbf{z} = [z_1, ...z_n]^{\mathrm{T}} \in \mathbb{R}^n$ where z_i represents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

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 SCM via the *linear layer* and then pass $\mathbf{z}$ into the *DAG layer* to reconstruct itself.

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 SCM 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 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 $\mathbf{A}$ [158]:

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

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

   $\mathcal{L}_{tanh} = 1/\text{tanh}(\frac{1}{n^2}\sum_{i=1}^{i=n}\sum_{j=1}^{j=n}|A_{i,j}| + \delta)$, 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:

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

Thus the overall loss function is $\mathcal{L} = -\mathcal{L}_{ELBO} + \lambda_1\mathcal{L}_{acyc} + \lambda_2\mathcal{L}_{tanh} + \lambda_3\mathcal{L}_{dag}$. 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.

	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.

3.7.3 ECG Reconstruction Performance of Personalized Digital Twins

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.

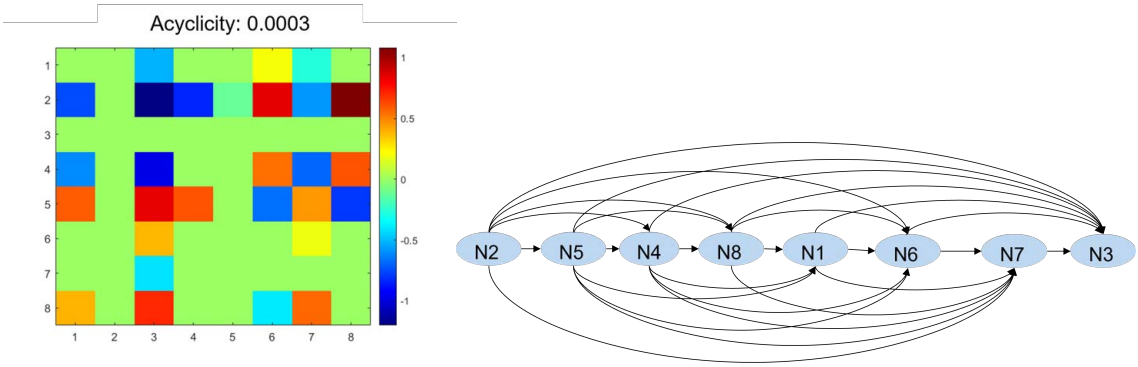
Quantitative Evaluation Metrics of Effect For Causal Analysis:

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 [57]. 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 [57]. 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 [57]. **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 [57]. 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.

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

	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).

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 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

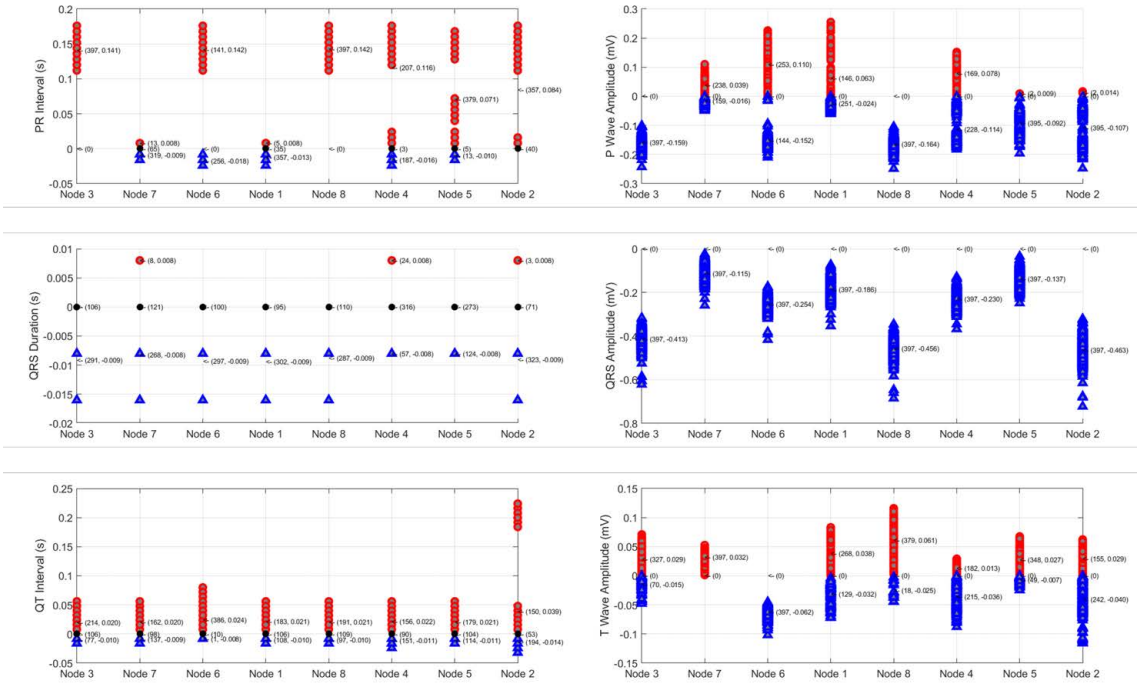


Figure 3.19: Distributions of the difference between the inferred ECG and the intervened ECGs for each evaluation metric, showing the impact of intervening each node in the latent causal representation. Red circle markers represent the increased value in the metrics after intervention and blue triangle 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.

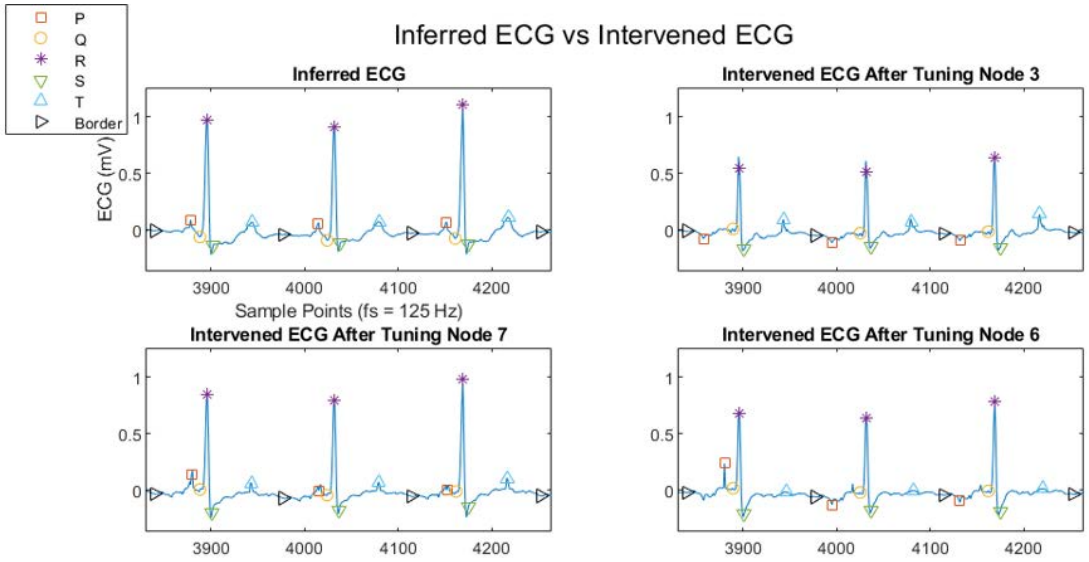


Figure 3.20: Visualization of the inferred ECG and intervened ECGs after tuning Nodes 3, 7, and 6, respectively. Established algorithms [104, 118, 119] are applied to detect the P, Q, R, S, and T fiducial points. The border point is defined as the 60%:40% segmentation point between each RR interval.

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 [39, 144] 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 assistance and professional input from doctors, we believe that a more clinically solid and well-rounded causal analysis can be conducted.

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

Chapter 4

A Multi-Channel Ratio-of-Ratios Method for Noncontact Hand Video Based SpO_2 Monitoring Using Smartphone Cameras

4.1 Related Works

4.1.1 Contact-based SpO_2 measurement using smart devices

It is of significance to realize early detection of changes in SpO_2 to facilitate timely management of asymptomatic patients with clinical deterioration. Conventional SpO_2 measurement methods rely on contact-based sensing, such as pulse oximetry introduced in Chapter 1.1.4 designed by the RoR principle.

With the ubiquity of smartphones and the growing market of smart fitness devices, the RoR principle has been applied to new nonclinical settings for SpO_2 measurement. Apple Watch Series 6 has blood oxygen measurement functionality, and it requires skin contact with the watch neither too tight nor too loose for the best results [64]. The recent scientific literature also explored methods for SpO_2 estimation using a smartphone. These methods require a user to use his/her fingertip to cover an optical sensor and a nearby light source to capture the reemitted light from the illuminated tissue [17, 37, 92, 117,

134]. In this setup, an adapted ratio-of-ratios model is utilized with the red and blue (or green) channels of color videos in lieu of the traditional narrowband red and infrared wavelengths.

The aforementioned SpO_2 estimation methods based on smartphones and smartwatches are contact-based. It can present the risk of cross-contamination between individuals using the same measurement device. An additional issue with contact-based methods is that they may irritate sensitive skin or a sense of burning from the heat built up if a fingertip is in contact with the flashlight for an extended period of time. Also, pulse oximeters may not be widely available in marginalized communities and some undeveloped countries [61].

4.1.2 Noncontact SpO_2 measurement using cameras

Researchers have recently investigated measuring the saturation of blood oxygen by means of contactless techniques [54, 77, 142, 143]. These methods typically acquire a user's face video under ambient light with CCD cameras to estimate SpO_2 from pulsatile information of monochromatic wavelengths. Shao et al. [123] also use a facial video-based method to monitor SpO_2 that is implemented using a CMOS camera with a light source alternating at two different wavelengths. Tsai et al. [140] acquire hand images with CCD cameras under two monochromatic lights to analyze SpO_2 from the reflective intensity of the shallow skin tissue. These contactless methods can provide alternatives to contact-based SpO_2 measurements for individuals with finger injuries or nail polish [31, 157], for whom the traditional pulse oximeters may be inaccurate. However, the setups

used in the abovementioned studies use either high-end monochromatic cameras with selected optical filters or controlled monochromatic light sources, making it expensive and not common for daily use.

As more economical camera devices, smartphones and webcams are also applied for contactless SpO_2 estimation. Most of the SpO_2 estimation works using digital RGB cameras under ambient light [12, 22, 113, 133] adapt the conventional RoR model based on the red and infrared wavelengths directly to the use of red and blue channels of RGB videos. It is worth noting that the SpO_2 data collected in [22, 113] only covers a small dynamic range (mostly above 95%), and Tarassenko *et al.* [133] and Bal *et al.* [12] show a fitted linear relation between RoR and SpO_2 for only several minutes of data. The limitations as mentioned above can be due to: i) Signals extracted from the red and blue channels are noisier than those extracted from the green channel [145], and ii) Unlike the narrowband signals being modeled in the conventional RoR model, the RGB color channels capture a wide range of wavelengths from the ambient light. The aggregation of the broad range of wavelengths lowers the optical difference between Hb and HbO_2 and makes it less optically selective than narrowband oximeter sensors and more challenging for SpO_2 sensing. So we are motivated to disentangle the aggregation effect through a meaningful combination of the pulsatile signals from all three channels of RGB videos to distill the SpO_2 information.

4.2 Ratio-of-ratios (RoR) Model for Noncontact SpO_2 Measurement

Consider a light source with the spectral distribution $I(\lambda)$ illuminating the skin and a remote color camera with spectral responsivity $r(\lambda)$ recording an image. According to the skin-reflection model [149], the color camera will receive the specularly reflected light from the skin surface and the diffusely reflected light from the tissue-light interaction that contains the pulsatile information. Based on the assumption proposed in [54] that the specular reflection can be ignored if the color change from movement is properly treated and minimized, the camera sensor response at time t can be expressed as:

$$\mathcal{S}_c(t) = \int_{\Lambda_c} I(\lambda) \cdot e^{-\mu_d(\lambda,t)} \cdot r_c(\lambda)\, d\lambda. \tag{4.1}$$

where the λ is the wavelength. The integral range Λ_c is the sensitive response wavelength band of the c th channel of the camera, $I(\lambda)$ is the spectral intensity of the light source, $\mu_d(\lambda, t)$ is the diffusion coefficient, and $r_c(\lambda)$ is the sensor response of the c th channel of the camera.

According to Beer-Lambert's law, the diffusion coefficient $\mu_d(\lambda, t)$ can be expanded into:

$$\mu_d(\lambda, t) = \varepsilon_t(\lambda) C_t l_t + [\varepsilon_{\text{Hb}}(\lambda) C_{\text{Hb}} + \varepsilon_{\text{HbO}_2}(\lambda) C_{\text{HbO}_2}] \cdot l(t). \tag{4.2}$$

where ε_{Hb}, $\varepsilon_{\text{HbO}_2}$, and ε_t are the extinction coefficients of arterial deoxyhemoglobin, arterial oxyhemoglobin, and other tissues including the venous blood vessel, respectively. C_t, C_{Hb}, and C_{HbO_2} are the concentration of the corresponding substances. l_t is the path length that the light travels in the tissue, which is assumed to be static and invariant of

time. $l(t)$ is the path length that the light travels in the arterial blood vessels. It is modeled as time-varying because the arteries will dilate with increased blood during systole compared to diastole.

When the camera is monochromatic, incoming light is filtered by a narrowband optical filter, or the light source is a narrowband LED, the integral range Λ_c can be simplified to a single value λ_i, such that the response of the camera sensor in Eq. (4.1) can be written as:

$$\mathcal{S}_c(t) = I(\lambda_i) \cdot e^{-\varepsilon_t(\lambda_i) C_t l_t} \cdot r_c(\lambda_i) \cdot e^{-[\varepsilon_{\text{Hb}}(\lambda_i) C_{\text{Hb}} + \varepsilon_{\text{HbO}_2}(\lambda_i) C_{\text{HbO}_2}] \cdot l(t)}. \tag{4.3}$$

Let $\Delta l = l_{\text{max}} - l_{\text{min}}$ denote the difference of the light path of the pulsatile arterial blood between diastole when $l(t) = l_{\text{min}}$ and systole when $l(t) = l_{\text{max}}$. Then the ratio of the response of the c th channel of the camera sensor during diastole and systole is:

$$\text{R}(\lambda_i) = \log \left(\frac{\mathcal{S}_c|_{l=l_{\text{min}}}}{\mathcal{S}_c|_{l=l_{\text{max}}}} \right) \tag{4.4a}$$

$$= [\varepsilon_{\text{Hb}}(\lambda_i) C_{\text{Hb}} + \varepsilon_{\text{HbO}_2}(\lambda_i) C_{\text{HbO}_2}] \cdot \Delta l. \tag{4.4b}$$

The ratio-of-ratios (RoR) between two different wavelengths λ_1 and λ_2 is:

$$\text{RoR}(\lambda_1, \lambda_2) = \frac{R(\lambda_1)}{R(\lambda_2)} = \frac{\varepsilon_{\text{Hb}}(\lambda_1) C_{\text{Hb}} + \varepsilon_{\text{HbO}_2}(\lambda_1) C_{\text{HbO}_2}}{\varepsilon_{\text{Hb}}(\lambda_2) C_{\text{Hb}} + \varepsilon_{\text{HbO}_2}(\lambda_2) C_{\text{HbO}_2}}. \tag{4.5}$$

Since $\text{SpO}_2(\%) = \frac{C_{\text{HbO}_2}}{C_{\text{HbO}_2} + C_{\text{Hb}}}$, the relation between RoR and SpO_2 can be derived from

Eq. (4.5) as:

$$\mathrm{SpO_2} = \frac{\varepsilon_{\mathrm{Hb}}(\lambda_1) - \varepsilon_{\mathrm{Hb}}(\lambda_2)\cdot \mathrm{RoR}}{\varepsilon_{\mathrm{Hb}}(\lambda_1) - \varepsilon_{\mathrm{HbO_2}}(\lambda_1) + [\varepsilon_{\mathrm{HbO_2}}(\lambda_2) - \varepsilon_{\mathrm{Hb}}(\lambda_2)]\cdot \mathrm{RoR}} \tag{4.6a}$$

$$\approx \alpha \cdot \mathrm{RoR} + \beta. \tag{4.6b}$$

where the linear approximation can be obtained by Taylor expansion.

The linear RoR model in Eq. (4.6b) has been applied under different $\mathrm{SpO_2}$ measurement scenarios. For pulse oximeters, $\lambda_1 = 660$ nm and $\lambda_2 = 940$ nm are used to leverage the optical absorption difference of Hb and $\mathrm{HbO_2}$ at the two wavelengths. In some prior art using narrowband light sources or monochromatic camera sensors [77, 123] for contactless $\mathrm{SpO_2}$ monitoring, different combinations of (λ_1, λ_2) have been explored. In the prior art using consumer-grade RGB cameras [12, 22, 113, 130, 133], only two out of the three available RGB channels were used for the linear RoR model.

Among the abovementioned $\mathrm{SpO_2}$ estimation methods using consumer-grade RGB cameras, the $\mathrm{SpO_2}$ data collected in [22, 113] only cover a small dynamic range (mostly above 95%), which is not very meaningful. Bal et al. [12] and Tarassenko et al. [133] show a fitted linear relation between RoR and $\mathrm{SpO_2}$ for data that last for merely several minutes. These limitations can be attributed to that, unlike the signals captured in the narrowband setting that is modeled precisely by Eq. (4.3) and Eq. (4.4), all three RGB color channels capture a wide range of wavelengths from the ambient light, as is described in Eq. (4.1). The aggregation of the broad range of wavelengths lowers the optical difference between Hb and $\mathrm{HbO_2}$ and makes it less optically selective than narrowband sensors used in oximeters. To address this issue, we disentangle the aggregation through a careful

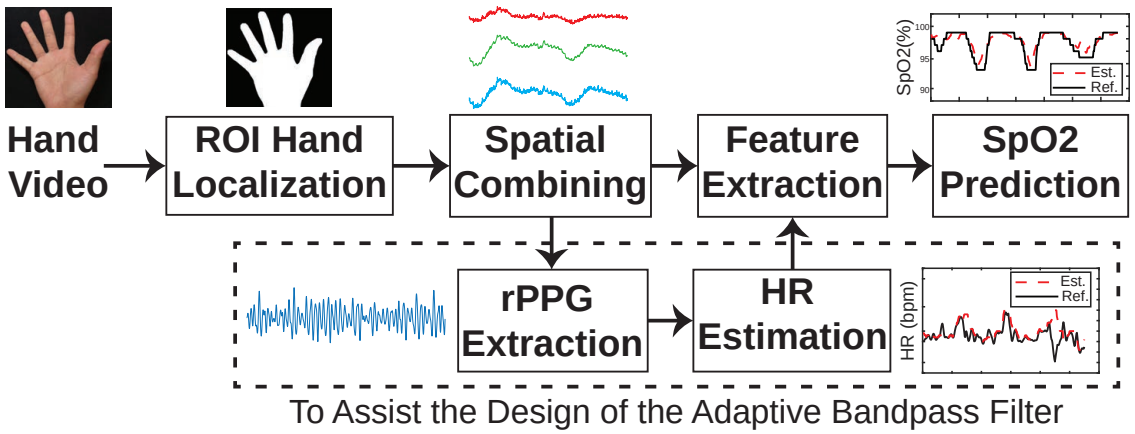


Figure 4.1: System illustration for the SpO_2 prediction using the smartphone captured hand videos. The pixels from the hand region are utilized for prediction, and an rPPG signal is extracted for heart rate (HR) estimation. Multi-channel RoR features are derived from the spatially combined RGB signals with the help of the HR-guided filters. The extracted features are then used for SpO_2 prediction.

combination of the pulsatile signals from all three channels of RGB videos to efficiently distill the SpO_2 information.

4.3 Proposed Multi-Channel RoR Method

In this work, we propose a multi-channel RoR method for noncontact SpO_2 monitoring using hand videos captured by smartphone cameras under ambient light. Fig. 4.1 illustrates the proposed procedure for the SpO_2 estimation from the smartphone captured hand videos. First, the hand is detected as the region of interest (ROI) for each frame. Second, the spatial average from the ROI is calculated to obtain three time-varying signals of RGB channels. The averaged RGB signals are extracted for two purposes: i) to estimate the heart rate (HR), and ii) to acquire the filtered cardio-related AC components using an HR-based adaptive bandpass filter. Third, the ratio between the AC and the DC components for each color channel and the pairwise ratios of the resulting three ratios are

computed as the features for a regression model where SpO_2 is treated as the label. The details of each step are provided as follows.

4.3.1 ROI Localization and Spatial Combining

First, we manually draw a rectangle to include the target hand region. This RGB region is converted to YCrCb color space, and the Cr channel is used [23] to determine a threshold that differentiates the skin pixels from the background based on the Otsu algorithm [103]. We apply an erosion and a dilation algorithm with a median filter to exclude noise pixels outside of the binary hand mask region. The final hand-shaped mask is considered as the ROI, and an example is shown in the second picture in Fig. 4.1. For all n frames in the video, we calculate the spatial average of the RGB channels in the ROI as $\mathbf{A} = [\bar{\mathbf{r}}; \bar{\mathbf{g}}; \bar{\mathbf{b}}]$, where $\bar{\mathbf{r}}, \bar{\mathbf{g}}, \bar{\mathbf{b}} \in \mathbb{R}^{1\times n}$, and $\mathbf{A} \in \mathbb{R}^{3\times n}$.

4.3.2 rPPG Extraction and HR Estimation

Typically in the RoR method, after the matrix $\mathbf{A}$ in Section 4.3.1 is calculated, the AC component for each channel of $\mathbf{A}$ is determined by either the standard deviation [117] or the peak-to-valley amplitude [123]. Since the signal-to-noise ratio (SNR) is lower for the video captured by a smartphone in a contactless manner, we propose to use an adaptive bandpass filter centered at the HR to filter the RGB channel signals and extract the AC components more precisely.

The HR can be measured contact-free by capturing the pulse-induced subtle color variations of the skin. The pulse signal, referred to as remote photoplethysmogram

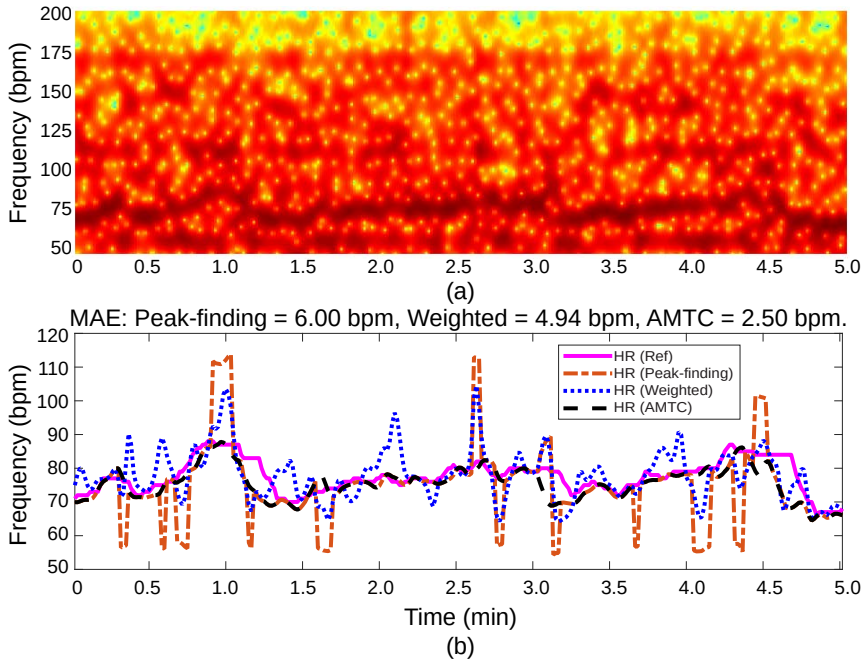


Figure 4.2: (a) Spectrogram of an rPPG signal. (b) Reference HR signals and HR signals estimated by the "naive" algorithm, the weighted energy frequency estimation algorithm, and the AMTC algorithm, respectively. The mean absolute error (MAE) of the HR estimation algorithms are 6.00 bpm, 4.94 bpm, and 2.50 bpm, respectively.

(rPPG), can be obtained by applying the plane-orthogonal-to-skin (POS) algorithm [149], which defines a plane orthogonal to the skin tone in the RGB space for robust rPPG extraction. The HR is then tracked from the rPPG signal via a state-of-the-art adaptive multi-trace carving (AMTC) [163] algorithm that tracks the HR from the spectrogram of rPPG by dynamic programming and adaptive trace compensation.

To study the role of accurate HR tracking for feature extraction, we also implemented a peak-finding method and a weighted energy method for frequency estimation [56] to compare with AMTC. The peak-finding method takes the peaks of the squared magnitude of the Fourier transform of rPPG as the estimated HR values, which was used in [142] and [54]. The weighted energy method finds the heart rate by weighing the frequency bins in the corresponding frame of the spectrogram of rPPG. Compared to the

peak-finding method, the weighted energy method is more robust to outliers in frequency. Fig. 4.2 illustrates an example of the HR estimation results by the peak-finding method, the weighted energy algorithm, and AMTC, respectively.

4.3.3 Feature Extraction

We use a processing window of 10 seconds with a step size of 1 second to segment the whole video into L windows. Within each window, the DC and AC components of the RGB channels are calculated to build a feature vector $\mathbf{f}$.

DC component We use a second-order lowpass Butterworth filter with a cutoff frequency of 0.1 Hz. The DC component is estimated using the median of the lowpass filtered signal of each window.

AC component The estimated heart rate values from Section 4.3.2 are used as the center frequencies for the adaptive bandpass (ABP) filters to extract the AC components of the RGB channels, which eliminates all frequency components that are unrelated to the cardiac pulse. We adopt an 8th-order Butterworth bandpass filter with ± 0.1 Hz (± 6 bpm) bandwidth, centering at the estimated HR of the current window. The magnitude of the AC component is estimated using the average peak-to-valley amplitudes of the filtered signals within the current processing window.

We define the normalized AC components at the ith window as $R(i,c) = \frac{\text{AC}(i,c)}{\text{DC}(i,c)}$, where $c \in \{r, g, b\}$ represents color channel and $i \in \{1, 2, ..., L\}$. We define the multi-channel ratio-of-ratios based feature vector of the ith window as

$\mathbf{f}_i = [R(i,r), R(i,g), R(i,b), \frac{R(i,r)}{R(i,g)}, \frac{R(i,r)}{R(i,b)}, \frac{R(i,g)}{R(i,b)}] \in \mathbb{R}^{1\times 6}$.

4.3.4 Regression and Postprocessing

As a proof-of-concept, we use linear regression and support-vector-regression (SVR) to learn the mapping between the features and the SpO_2 values.

The linear regression has limited learning capability since it captures only the linear relationship. So we use it as a baseline approach. In the objective function in Eq. (4.7), $\mathbf{y} = [y_1, ..., y_l] \in \mathbb{R}^{l\times 1}$ is the target SpO_2 value, $\omega \in \mathbb{R}^{6\times 1}$ is the predictor, and $\mathcal{F}$ is the feature matrix that serves as input. We add an L_2 regularization term in Eq. (4.7) to avoid collinearity. To select the optimal weight λ for the L_2 regularization term, we use 5-fold cross-validation.

$$\min_{\omega} \|\mathbf{y} - \mathcal{F}\omega\|_F^2 + \lambda \|\omega\|_2^2 . \tag{4.7}$$

SVR models are adopted for exploring possibly nonlinear relation between the feature vectors and the SpO_2 estimation. The Libsvm library [24] is used for training the "ϵ-SVR" in Eq. (4.8). In our implementation, we use the nonlinear Radial Basis Function (RBF) kernel for the SVR. The hyperparameters, including the penalty cost C, and the kernel parameter γ of kernel function $K(\mathbf{f}_i, \mathbf{f}_j) = \boldsymbol{\phi}(\mathbf{f}_i)^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{f}_j) = \exp\left(-\gamma \|\mathbf{f}_i - \mathbf{f}_j\|^2\right)$ are selected via grid search and a 5-fold cross-validation.

$$\begin{aligned} \min_{\omega,b,\xi,\xi^*} \frac{1}{2}\|\omega\|_2^2 + C\cdot\sum_{i=1}^{l}(\xi_i+\xi_i^*) \\ \text{s.t.} \quad \phi(f_i)\cdot\omega+b-y_i \leq \epsilon+\xi_i, \\ y_i-\phi(f_i)\cdot\omega-b \leq \epsilon+\xi_i^*, \\ \xi_i,\xi_i^* \geq 0, i=1,...,l. \end{aligned} \tag{4.8}$$

Once an estimated weight vector $\hat{\mathbf{w}}$ is learned from the linear or support vector regression, $\hat{\mathbf{w}}$ is then used to predict a preliminary SpO_2 signal. Finally, a 10-second moving average window is applied to smooth out the preliminarily predicted signal to obtain the final predicted SpO_2 signal.

4.4 Experimental Results

4.4.1 Data Collection

Fourteen volunteers, including eight females and six males, were enrolled in our study under protocol #1376735 approved by the University of Maryland Institutional Review Board (IRB), with age range between 21 and 30. Participants were asked to categorize their skin tone based on the Fitzpatrick skin types [10] shown in Fig. 4.3. There are two, eight, one, and three participants having skin types II, III, IV, and V, respectively. None of the participants had any known cardiovascular or respiratory diseases. During the data collection, participants were asked to hold their breath to induce a wide dynamic range of SpO_2 levels. The typical SpO_2 range for a healthy person is from 95% to 100%. By holding breath, the SpO_2 level can drop below 90%. Once the participant resumes

normal breathing, the SpO_2 will return to the level before the breath-holding.

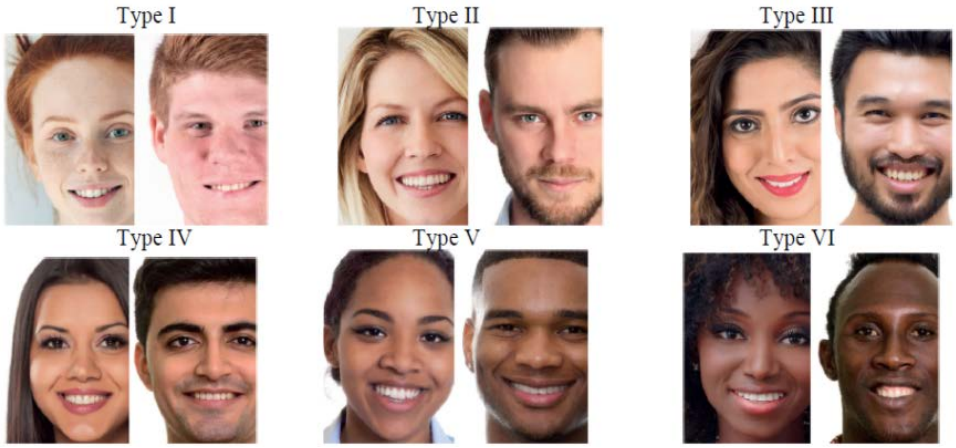


Figure 4.3: Fitzpatrick skin types [10].

Each participant was recorded for two sessions. During the recording, the participant sat comfortably in an upright position and put both hands on a clean dark foam sheet placed on a table. As shown in Fig. 4.4, the palm side of the right hand and the back side of the left hand were facing the camera. These two hand-video capturing positions are defined as *palm up (PU)* and *palm down (PD)*, respectively. The participant was asked to place his/her hands still on the table to avoid hand motion. Simultaneously, a Contec CMS-50E pulse oximeter was clipped to the left index finger to measure the participant's SpO_2 level at a sampling rate of 1Hz. As we have reviewed earlier, the oximeter is adopted clinically to be within a $\pm 2\%$ deviation from the invasive, gold standard for SpO_2 [108], so we use the oximeter measurement results as the reference in our experiments. An iPhone 7 Plus camera was fixed by a smartphone stand mounted on a tripod for video recording at a sampling rate of 30 fps. The video started 30 seconds before the oximeter started and stopped immediately after the oximeter ended to allow for proper time synchronization. The participants were asked to hold their breath for generally 30–40 seconds

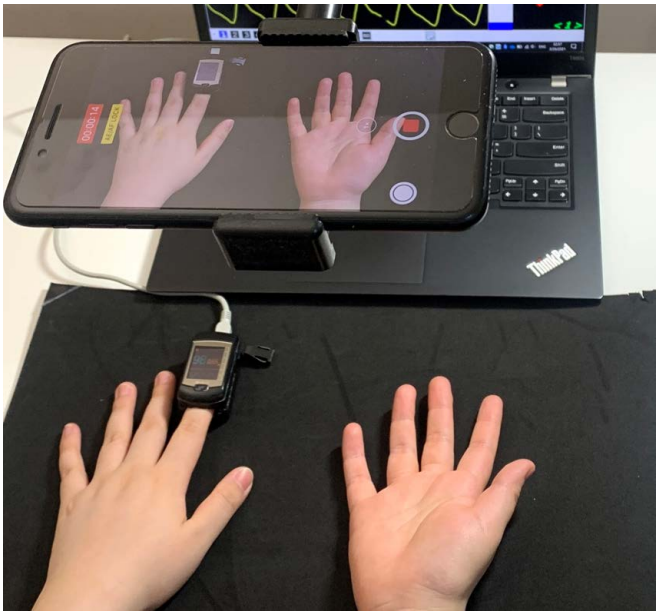


Figure 4.4: Experimental setup for data collection of hand videos and reference signals using an oximeter. The left index finger was placed in a CMS-50E pulse oximeter to record the reference HR and SpO_2 signals. The smartphone camera is recording the video of both hands.

to lower the SpO_2 level, as long as they were comfortable and able to do so. Then the participants resumed normal breathing for generally 30–40 seconds until they recovered and felt ready for the next breath-holding. The recovery period was long enough for the participants' SpO_2 to return to the levels before the breath-holding. The aforementioned process is defined as one breath-holding cycle. In each session, the breath-holding cycles were repeated three times. After the first session, the participants were asked to relax for at least 15 minutes before attending the second session for data collection. From our data collection protocol using breath-holding, we were able to obtain the SpO_2 measurements ranging from 89% to 99%.

The total length of recording time for all fourteen participants is 138.9 minutes. In terms of each participant, the minimum duration is 103 seconds and the maximum duration is 468 seconds. The average duration is 298 seconds. The current data size is

relatively small for large-scale neural network training. This is by a large part due to the restrictions for human subject related data collection imposed during the COVID-19. The available data, however, is adequate for our principled multi-channel signal based approach to SpO_2 monitoring, showing a benefit of combining signal processing and biomedical knowledge and modeling with data than the primarily data-driven approach.

Delay Estimation of Pulse Oximeter: When the CMS-50E oximeter is turned on and ready for measurement, the first reading is displayed a few seconds after the finger is inserted. This delay may be due to the oximeter's internal firmware startup and algorithmic processing. Since we need to synchronize the video and the oximeter readings using their precise starting time stamps, the delay in the oximeter can introduce misalignment errors in the reference data that we use to train the regression model. To avoid misalignment, we first estimate the delay and then subtract it from the oximeter's internal timestamp as the corrected oximeter's timestamp. To estimate the internal delay, we asked one participant to repeatedly place the left index finger, middle finger, and ring finger into the oximeter 50 times each and obtained the average delay time of 1.8s, 1.9s, and 1.7s, respectively. Because the left index finger is used for reference data collection in our setup, we take 1.8s as the delay. To further examine whether there exists any difference among the delays from the three fingers, we conducted a one-way ANOVA test. The p-value is 0.14, which shows no statistically significant different delays among the three fingers.

4.4.2 Performance Metrics

The performance of the algorithm is evaluated using the mean absolute error (MAE) and Pearson's correlation coefficient ρ given in (4.9). Note that the correlation is adopted to evaluate how well the trend of SpO_2 is tracked.

$$\mathrm{MAE}(\mathbf{y},\hat{\mathbf{y}}) = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|, \quad \rho(\mathbf{y},\hat{\mathbf{y}}) = \frac{(\mathbf{y}-\bar{y})^{\mathrm{T}}(\hat{\mathbf{y}}-\bar{\hat{y}})}{\|\mathbf{y}-\bar{y}\|_2 \left\|\hat{\mathbf{y}}-\bar{\hat{y}}\right\|_2}. \tag{4.9}$$

where $\mathbf{y} = [y_1, ..., y_N]^{\mathrm{T}}$, $\hat{\mathbf{y}} = [\hat{y}_1, ..., \hat{y}_N]^{\mathrm{T}}$, $\bar{y}$, and $\bar{\hat{y}}$ denote the reference SpO_2 signal, the estimated SpO_2 signal, the average values of all coordinates of vectors $\mathbf{y}$ and $\hat{\mathbf{y}}$, respectively. We adopt the correlation metric to evaluate how well the trend of the SpO_2 signal is tracked.

4.4.3 Results From Proposed Algorithm

In this subsection, we use the training data from one participant to train the regression model for the prediction of his/her testing session recorded later. We call the aforementioned training and testing procedure the *participant-specific* mode in which the models are specifically learned for each participant. We will discuss the *leave-one-out* mode of the performance of the proposed algorithm in Section 4.4.5.

Fig. 4.5 presents the learning results for all the participants using SVR for PU cases. Both training and testing sessions are shown for each participant. The SpO_2 curves in each session contain three dips that are resulted from breath-holding, except for participant #8 who had a shorter session due to limited tolerance of breath-holding. For each participant,

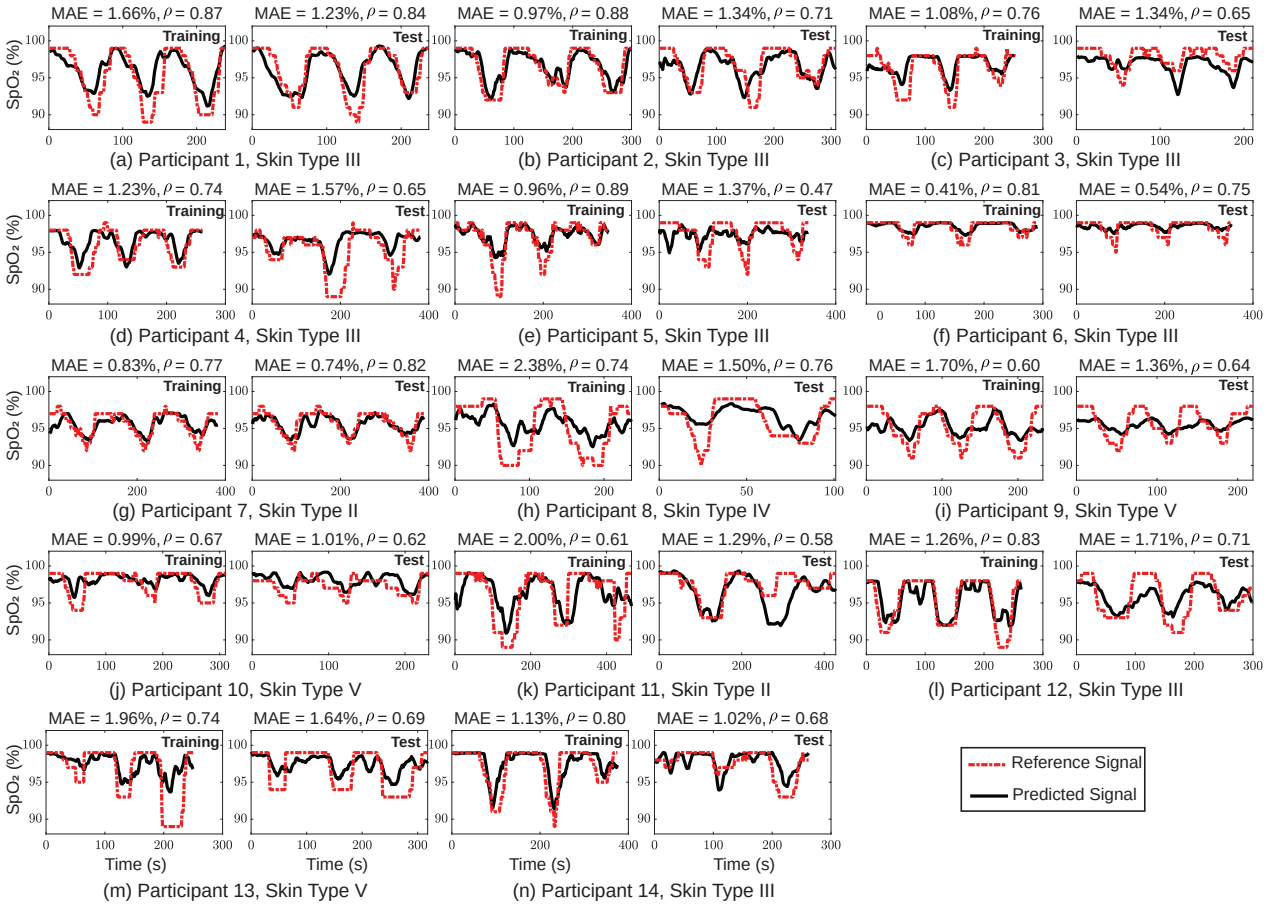


Figure 4.5: Predicted SpO_2 signals for all participants using SVR when the palm is facing the camera, i.e., the palm-up scenario. Prediction results of training and testing sessions are shown for each participant with reference SpO_2 in red dash lines and predicted SpO_2 in solid black lines. The higher the correlation ρ and the lower the MAE, the better the predicted SpO_2 captures the trend of the reference signal.

		Training		Testing	
		MAE	Correlation ρ	MAE	Correlation ρ
LR	PU	1.69% (±0.57%)	0.63 (±0.16)	1.52% (±0.54%)	0.62 (±0.11)
	PD	1.74% (±0.76%)	0.61 (±0.21)	1.53% (±0.53%)	0.56 (±0.21)
SVR	PU	**1.33%** (±0.54%)	**0.76** (±0.09)	**1.26%** (±0.33%)	**0.68** (±0.10)
	PD	1.35% (±0.45%)	0.75 (±0.09)	1.28% (±0.40%)	0.65 (±0.14)

Table 4.1: Performance of the proposed method. Results using linear regression (LR) and support vector regression (SVR) for both sides of the hand are quantified in terms of the sample mean and sample standard deviation (in parentheses).

we provide the skin-tone information in the subplot and show the accuracy indicators, MAE and ρ, for SpO_2 prediction. In all training sessions, MAE is below 2.4% and ρ is above 0.6. From this observation, we find that all the predicted SpO_2 signals in the training sessions are closely following the reference signals' trends, despite the exact value differences between the predicted and the reference signals, such as the differences around the last dip for participant #13. Furthermore, all testing MAE values are within 1.8%, suggesting that those trained models adapt well to the testing data. While there are a few cycles where the predicted signal does not fully follow the reference signal, such as the second dip for participant #4 and participant #11, the trends are consistent.

Table 4.1 summarizes the training and testing SpO_2 estimation performance of both LR and SVR based methods for both PU and PD cases. The best performance is achieved using the SVR method in the PU case. We further examine the difference between the two regression methods using boxplots in Fig. 4.6(a) that show the distributions of the correlation ρ for testing by LR and SVR, respectively. Each boxplot in Fig. 4.6(a) contains both PU and PD cases from all participants. The results are compared in terms of the

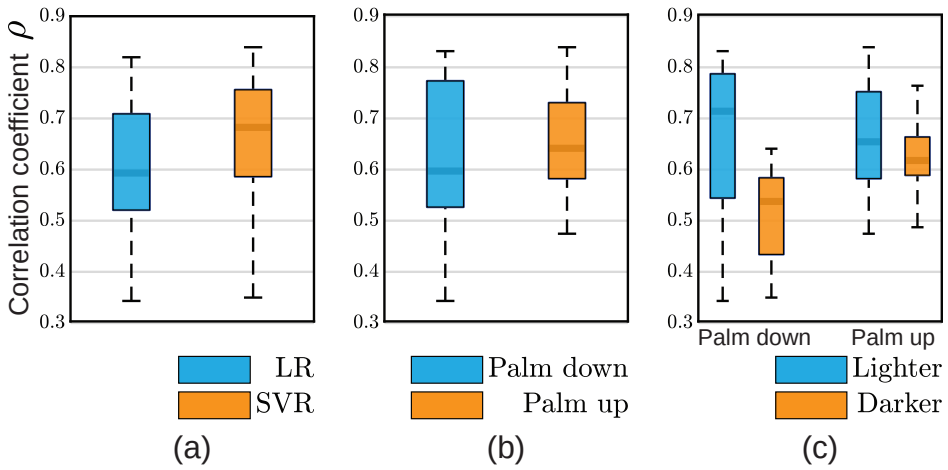


Figure 4.6: Boxplots of testing correlation coefficient ρ for all participants when grouped using different criteria. (a) Distributions contrasting linear and support vector regressions. (b) Distributions of palm-up and palm-down cases. (c) A detailed breakdown of (b) in terms of skin-tone subgroups.

median and the interquartile range (IQR). IQR quantifies the spread of the distribution by measuring the difference between the first quartile and the third quartile. The boxplots in Fig. 4.6(a) reveal that the SVR method outperforms LR with a higher median of 0.68 compared to 0.59 and with a narrower IQR of 0.17 compared to 0.19. This suggests that there may exist a nonlinear relationship between the extracted features and the SpO_2 values.

To examine the impact of the side of a hand and the skin tone on the performance of SpO_2 estimation, we analyze the following two research questions: (i) whether the side of the hand makes a difference in lighter skin (type II and III) or darker skin (type IV and V) or mixed skins (all participants), and (ii) whether the different skin tones matter in PU or PD case.

To answer question (i), we first focus on the distributions from PU and PD cases in Fig. 4.6(b) with each boxplot representing the correlation ρ in testing for all participants.

We observe that the PU case outperforms the PD case with a higher median of 0.64 compared to 0.60 and a narrower IQR of 0.15 compared to 0.25. We then zoom into each subgroup of skin tones shown in Fig. 4.6(c). For the lighter skin group, even though the median of PD case is 0.71, which is 9% better than that of PU, the IQR of PD case is 0.24, which is worse than the IQR of 0.17 of PU case. This suggests that the distributions are comparable between PU and PD cases for the lighter skin group. For the darker skin group, the PU case outperforms the PD case with a higher median of 0.62 compared to 0.54 and a narrower IQR of 0.07 compared to 0.15. In summary, there is no substantial difference between PU and PD cases in the lighter skin group, whereas, for the darker skin group and overall participants, the PU case is better than the PD case.

To answer question (ii), we first focus on the left two boxplots of Fig. 4.6(c). In the PD case, the median of the lighter skin group is significantly larger than that of the darker skin group by 31%, however, the lighter skin group also has a larger IQR. This makes it difficult to make a conclusion from the median–IQR analysis, hence we apply the t-test to complement our analysis. We note that the p-value is $0.037 < 0.05$, showing that there is a significant difference between these two groups. In the PU case shown in the right half of Fig. 4.6(c), the medians of the lighter skin group and darker skin group are 0.65 and 0.62, with IQR being 0.17 and 0.07, respectively. Thus, in our current dataset, no substantial performance difference is observed between lighter and darker skin tones in the PU case.

Method Index	Configuration		
	Multi-channel RoR features?	Narrow ABP filter?	Accurate HR tracking?
I	Two-channel RoR	✓	✓ (AMTC)
II	✓	No ABP	n/a
III	✓	Wide ABP	✓ (AMTC)
IV	✓	✓	Peak-finding
V	✓	✓	Weighted energy
Proposed	✓	✓	✓ (AMTC)

Table 4.2: Configurations for the ablation study of the proposed pipeline. The controlled experiments are conducted by replacing or removing one component at a time.

4.4.4 Ablation Study of Proposed Pipeline

In Sections 4.3.2 and 4.3.3, we have proposed three key designs in our algorithm, including a) the feature vector $\mathbf{f}$ containing pulsatile information from all RGB channels, b) the narrow ABP filter, and c) the passband of the ABP filter centered at precise HR frequency tracked by AMTC. To study the importance of each component, we conducted three controlled experiments by removing one factor at a time and the configurations of methods corresponding to the experiments are listed in Table 4.2. The results for the methods are illustrated in Fig. 4.7. The height of each bin shows the average correlation coefficient ρ or the MAE of SpO_2 estimation results from testing sessions (SVR, PU case) of all participants. Each pair of error bars 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 participants.

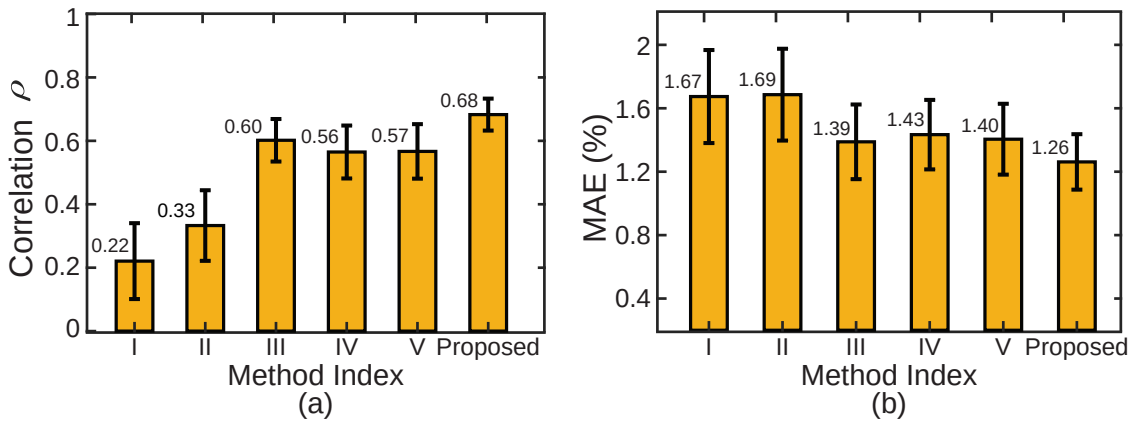


Figure 4.7: Ablation study of the proposed method. The bar plots are from testing sessions (SVR, PU case) of all participants. The error bars correspond to the 95% confidence intervals.

4.4.4.1 Advantage of The Proposed Multi-Channel RoR Over Two-Channel RoR

In this part, we compare our proposed algorithm with "**Method (I): RoR with nABP (AMTC)**." Method (I) follows the feature extraction method proposed in Section 4.3.3, including the *narrow adaptive bandpass filter (nABP)* centered at AMTC-tracked HR. The only exception is that, instead of using the feature vector $\mathbf{f}$ that contains multi-channel information, only the ratio of ratios between the red and blue channels as in traditional RoR methods is used.

Fig. 4.7 reveals that our proposed method outperforms method (I) by a big margin. More specifically, our proposed method improves the correlation coefficient from 0.22 to 0.68 and the MAE from 1.67% to 1.26%. This improvement confirms that our proposed multi-channel feature set helps with more accurate SpO_2 monitoring.

4.4.4.2 Contribution of Narrowband ABP Filter for Feature Extraction

Here we compare the following two methods to show the necessity of using a narrowband HR-guided bandpass filter:

- **Method (II): Feature vector without ABP** uses a nonadaptive, generic bandpass filter with the passband over $[1, 2]$ Hz, covering the normal range of heart rate in sedentary mode to replace the HR-based narrow ABP filter proposed in Section 4.3.3 for feature extraction.

- **Method (III): Feature vector with wide ABP (AMTC)** applies a wider ABP filter with ± 0.5 Hz bandwidth than the ± 0.1 Hz one used in our proposed method. This wider ABP filter's center frequency is provided by the AMTC tracking algorithm of the HR described in Section 4.3.2.

The bandpass filters used for methods (II) and (III) have the same bandwidth, 1 Hz. In terms of center frequency, method (II) used a fixed setting at 1.5 Hz, while method (III) is adaptively centered at the estimated HR value. Compared to method (II), method (III) has an improved testing MAE by 18%. Furthermore, compared to method (III), our proposed method with a narrow ABP filter improves the correlation coefficient ρ for testing by 13% and MAE by 9%, suggesting the contribution of the narrow HR-based ABP filter strategy for AC computation.

4.4.4.3 Importance of Accurate HR Tracking on SpO_2 Monitoring

We consider the following two methods to compare with our proposed method:

- **Method (IV): Feature vector with narrow ABP (peak-finding)** applies a narrow ABP filter of bandwidth ± 0.1 Hz for extracting the feature vector $\mathbf{f}$. The center frequency of the ABP filter is the HR estimated from the peak-finding algorithm described in Section 4.3.2.

- **Method (V): Feature vector with narrow ABP (weighted)** is similar to method (IV), except that the frequency estimation algorithm is replaced by the weighted energy in Section 4.3.2.

The averaged MAE of the HR estimation for all participants by the peak-finding algorithm, weighted frequency estimation algorithm, and AMTC algorithm are 7.11 (± 3.66) bpm, 6.42 (± 3.02) bpm, and 4.14 (± 1.72) bpm, respectively.

Fig. 4.7 shows that methods (IV) and (V) perform similarly with 0.56 vs. 0.57 for correlation ρ and 1.43% vs. 1.40% for MAE, respectively. Our proposed method guided by the AMTC tracked HR outperforms methods (IV) and (V) by 21% and 19% in correlation, and by 12% and 10% in MAE, respectively. These results suggest that the accurate HR estimation for ABP filter design improves the quality of the AC magnitude by preserving the most cardiac-related signal from RGB channels, which in turn helps with accurate SpO_2 monitoring.

4.4.5 Leave-One-Out Experiments

As a proof of concept and considering the currently limited amount of available data, we have so far discussed the SpO_2 estimation under the *participant-specific (PS)* scenario in Section 4.4 where the models are calibrated for each individual. This PS

	LOPartO		LOSessO		PS	
	MAE	ρ	MAE	ρ	MAE	ρ
PU	1.70% (±0.60%)	0.53 (±0.38)	1.59% (±0.58%)	0.55 (±0.36)	1.26% (±0.33%)	0.68 (±0.10)
PD	1.76% (±0.59%)	0.48 (±0.38)	1.70% (±0.59%)	0.50 (±0.39)	1.28% (±0.40%)	0.65 (±0.14)

Table 4.3: Testing results of leave-one-participant-out (LOPartO) and leave-one-session-out (LOSessO) experiments, measured in the sample mean and the sample standard deviation (in parentheses).

mode corresponds well to the trending "precision telehealth" that tailors the healthcare service to individuals.

In this subsection, we consider a more practical scenario where the test participant's data are never seen or only form a limited portion of the training data. In this scenario, we can develop a group-based model based on skin tone or other determinants of health, and for each subgroup, the model is "universal" and participant-independent. We will examine this group-based model through the following two modes of leave-one-out experiments:

- *Leave-one-session-out (LOSessO)*: when testing on a given participant, we include his/her training session data together with other participants' data for training.
- *Leave-one-participant-out (LOPartO)*: when testing on a given participant, we only use other people's data for training and leave out the data from this test participant.

We group the participants by skin type into lighter skin color (skin types II and III) and darker skin color (skin types IV and V) groups. We conduct LOSessO and LOPartO experiments on each subgroup and obtain the SVR generated testing results from all participants in Table 4.3. The MAE and correlation coefficient ρ improve from LOPartO to LOSessO to PS for both PU and PD cases. This result suggests that the precision

telehealth inspired PS mode is the most accurate approach to monitoring SpO_2 for an individual. Based on the overall results shown in Table 4.3, most participants demonstrate a consistent trend of the accuracy of SpO_2 estimation from LOPartO to LOSessO to PS case. The correlation ρ of participant #12 is less than -0.5 in both leave-one-out modes, suggesting that this participant may have some uncommon relation compared to others between the extracted features and SpO_2 values.

4.5 Discussions

4.5.1 Performance on Contact SpO_2 Monitoring

In addition to contact-free SpO_2 monitoring, we evaluate whether our proposed algorithm can be applied to a contact-based smartphone setup. To collect data, the left index finger covers the smartphone's illuminating flashlight and the nearby built-in camera, and the camera captures a pulse video at the fingertip. Another smartphone is used to simultaneously record a top view video of the back side of the right hand whose index finger is placed in the oximeter for SpO_2 reference data collection. One participant took part in this extended experiment where one training session with three breath-holding cycles was recorded, and three testing sessions were recorded 30 minutes after the training session.

In Table 4.4, we compare the performance of our proposed algorithm in both the contact-based and contact-free SpO_2 measurement settings. The conventional RoR models used in [117] and [92] were implemented as baseline models for contact-based SpO_2 measurement. In [117], the mean and standard deviation of each window from the red and blue channels are calculated as the DC and AC components. A linear model was built

		Training		Testing	
		MAE	ρ	MAE	ρ
Contact	RoR [117] (LR)	1.60%	0.54	1.38%	0.64
	RoR [117] (SVR)	1.14%	0.73	1.32%	0.60
	RoR [92] (LR)	1.47%	0.62	1.39%	0.63
	RoR [92] (SVR)	0.99%	0.83	1.27%	0.66
	Proposed	**0.91%**	**0.84**	**1.17%**	**0.81**
Contact-free	RoR (2-channel)	1.61%	**0.73**	1.75%	0.36
	Proposed	**1.36%**	0.62	**1.29%**	**0.68**

Table 4.4: Comparison of the proposed algorithm in both contact and contact-free SpO_2 estimation settings. The testing results are measured in the average MAE and correlation coefficient ρ.

to relate the ratio-of-ratios from the two color channels with SpO_2. In [92], the median of the pulsatile peak-to-valley amplitude is regarded as the AC component. For the two RoR methods, we implemented both LR and SVR. For contact-free SpO_2 measurement, we take the traditional two-color channel RoR method implemented in Section 4.4.4 as the baseline to compare with the proposed method.

Table 4.4 reveals that our proposed algorithm outperforms other conventional RoR models in contact-based SpO_2 monitoring. Even in the contact-free case, our algorithm presents a comparable performance to that of the contact-based cases, despite that the SNR of the fingertip video is better than the SNR from a remote hand video.

4.5.2 Resilience Against Blurring

In this subsection, we explore the robustness of our algorithm to the blurring effect on hand images. In the current setup, the hands are placed on a stable table with a cellphone camera acquiring the skin color of both hands. Ideal laboratory conditions are often not satisfied under practical scenarios, and the hand images captured by the

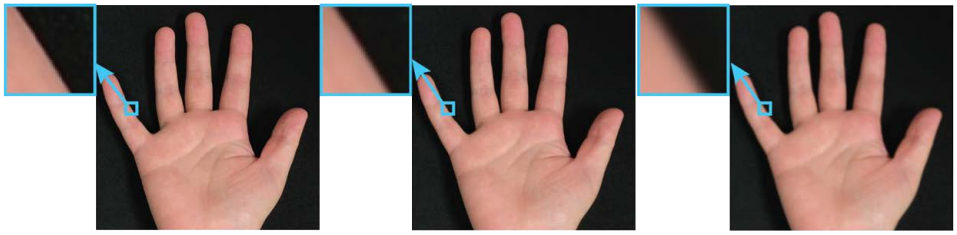


Figure 4.8: Illustration of blurring effects using different blurry levels σ on hand videos. The wider the kernel is, the blurrier the videos are.

	Training		Testing	
	MAE	ρ	MAE	ρ
$\sigma = 2.6$ blur (15×15 pixels)	1.41% ($\pm$0.50%)	0.72 ($\pm$0.11)	1.31% ($\pm$0.35%)	0.67 ($\pm$0.09)
$\sigma = 1.1$ blur (5×5 pixels)	1.42% ($\pm$0.59%)	0.70 ($\pm$0.16)	1.34% ($\pm$0.41%)	0.68 ($\pm$0.10)
No blur	1.33% ($\pm$0.54%)	0.76 ($\pm$0.09)	1.26% ($\pm$0.33%)	0.68 ($\pm$0.10)

Table 4.5: Simulation for Gaussian blurring effect on hand videos. SVR generated results for PU cases are listed for different σ and Gaussian kernel sizes. The results are quantified in terms of the sample mean and sample standard deviation (in parentheses).

cellphone cameras may be blurred due to being out of focus. The point spread function is modeled as a 2D homogeneous Gaussian kernel. The finite support of the kernel is defined manually to generate perceptually different blurry effects and then the standard deviation σ is computed based on the given support. To test different blurry effects, we experimented with two different blurry levels $\sigma = 1.1$ (5×5 pixels) and $\sigma = 2.6$ (15×15 pixels), respectively. We show the blurring effects in Fig. 4.8.

Table 4.5 presents the SVR generated results for PU cases with different σ and kernel sizes. We use the SVR, PU scenario to showcase here as it achieves the best SpO_2 prediction performance, which is verified in Section 4.4.3. From the table, we find that our algorithm is robust to the Gaussian blurring effect. After the $\sigma = 1.1$ blurring, the

testing ρ remains the same, and testing MAE is 6.3% higher than the no blurring case. After the $\sigma = 2.6$ blurring, the testing ρ is 1.5% lower and MAE is 4.0% higher than the no blurring case.

4.5.3 Limitations and Further Verification with Intermittent Hypoxia Protocols

From the recordings of our data collection protocol for voluntary breath-holding, we observed that HR and SpO_2 are correlated for many participants. That is, in one breath-holding cycle, when the participant starts to hold their breath, his/her HR increases and SpO_2 drops as the oxygen runs out. As he/she resumes normal breathing, his/her HR and SpO_2 recover to be within the normal range. Due to individuals' different physical conditions, in some participants, the peak of the HR signal and valley of the SpO_2 signal happen in such a short time interval that HR and SpO_2 are significantly negatively correlated. This observation is in line with the biological literature [53]. In the literature, breath-holding exercises were found to be able to yield significant changes in the cardiovascular system. In the central circulation, they caused significant changes in heart rate, and in the peripheral circulation, they caused significant changes in arterial blood flow and oxygen saturation.

Based on the above observation that HR is correlated with SpO_2 during breath-holding, we are curious whether our method also works for a different protocol where the instant HR change is relatively less correlated to SpO_2. An *intermittent hypoxia (IH) protocol* used in the literature shows that by receiving hypoxic air (inspired fraction of

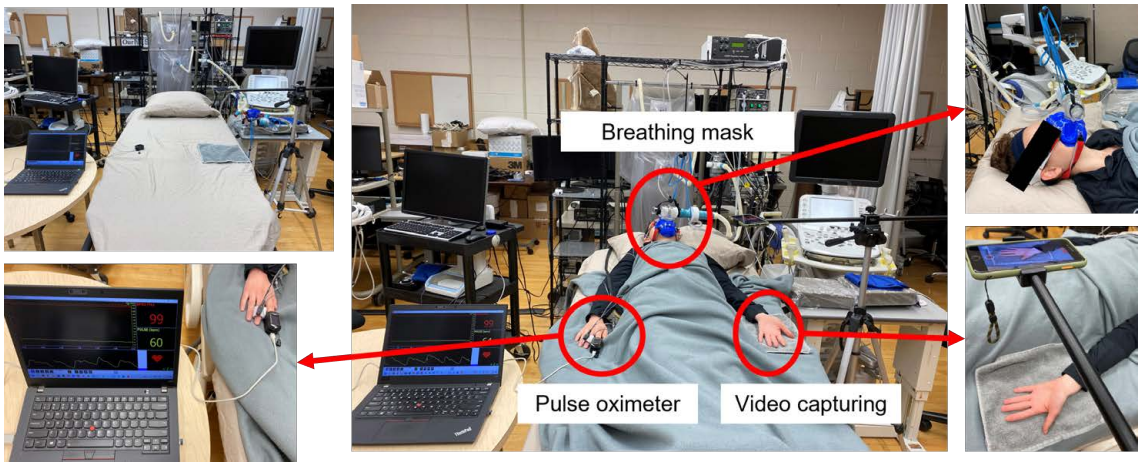


Figure 4.9: Experimental setup for the intermittent hypoxia protocol. The participant lies down on a bed with a mask controlling the breathing-in air which alternates between hypoxia and normoxia. The right index finger is clipped by the CMS-50E pulse oximeter to record the reference SpO_2 and HR signals. The palm side of the left hand (PU) is facing toward the smartphone camera during hand video recording sessions.

oxygen between 12% and 15%) intermittently with normoxic air, the participant can have a much milder HR change than breath-holding, while a significant decrease in SpO_2 can be achieved during the hypoxia [43]. The research restriction affecting human subject research in many U.S. institutions limited our ability to carry out the abovementioned hypoxia protocol before and as the restriction is eased recently, we investigate the performance of our proposed algorithm when applied to the new hypoxia protocol.

IH Protocol and Data Collection Setup:

Similar to the breath-holding protocol used in Chapter 4.4.1, the data collection setup of the IH protocol (shown in Fig. 4.9) includes a Contec CMS-50E pulse oximeter attached to the right index finger to measure the participant's SpO_2 and HR level as the reference and an iPhone 7 Plus camera mounted on a tripod for hand video recording. In lieu of holding breath to induce variation in SpO_2 values, in the IH protocol, the participant is equipped with a face mask that controls the breathing-in air. The face mask is

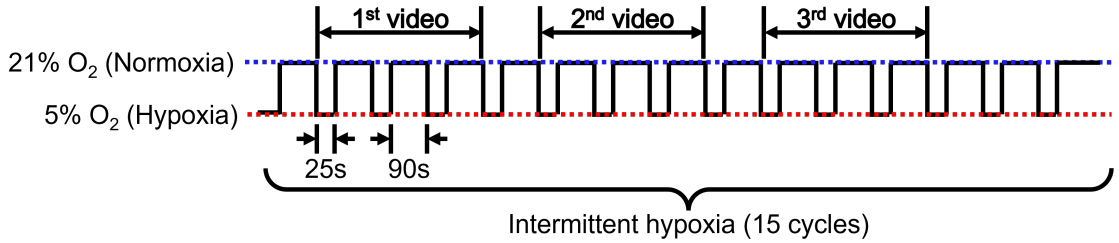


Figure 4.10: Illustration for the intermittent hypoxia (IH) protocol. The IH breathing is composed of 15 cycles of exposures to the alternating 25-second hypoxia (5% oxygen) period and the 90-second normoxia (21% oxygen) period. Three hand video sessions are recorded during the process and each video takes around 4.5 minutes. The first, second, and third videos start at the 2nd, 6th, and 10th IH cycle, respectively. In the practical data collection, the start time of the second and third video can be delayed by 1 or 2 cycles for the participants to adjust their hand positions after the previous video session.

connected to a one-way non-rebreathing valve, which is attached to a two-way switching valve. The two-way switching valve is used to control the input of either hypoxic air (5% oxygen, 3% carbon dioxide, balanced nitrogen) or room air (normoxia: 21% oxygen). Throughout the protocol, a switching valve is alternated between the acute (25-second) exposures to hypoxic air and the 90-second exposure to the normoxic medical gas for a total of 15 hypoxic events. Three hand video sessions are recorded for each participant during the process and each video takes around 4.5 minutes. The illustration for the procedure is shown in Fig. 4.10.

Overview of Participant Information and Collected Data:

Three participants, including one male and two females, were enrolled in the study under protocol #1511266 approved by the University of Maryland IRB, with one female's Fitzpatrick skin type being type I and the other two participants' being type II. One frame from the hand videos where the palm side facing the camera (PU case) of each participant is shown in Fig. 4.11. According to the IH protocol described in the previous paragraph, each participant had three hand video sessions recorded while their SpO_2 and HR were

(a) IH-1 (b) IH-2 (c) IH-3

Figure 4.11: Hand images from (a) a male participant whose Fitzpatrick skin type is II, (b) a female participant whose Fitzpatrick skin type is I, and (c) a female participant whose Fitzpatrick skin type is II.

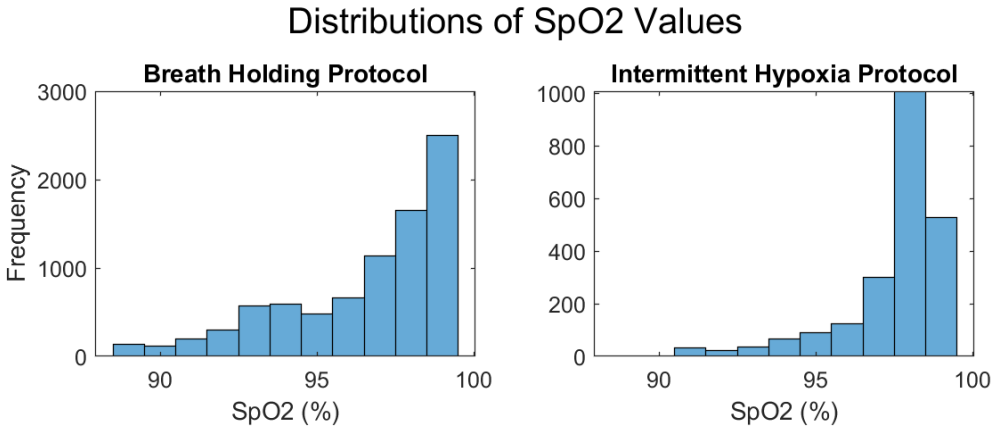


Figure 4.12: Comparison of the distributions of SpO_2 collected using the breath-holding protocol and the intermittent hypoxia protocol.

measured by the pulse oximeter during the intermittent hypoxia process. The histograms of SpO_2 values in the collected datasets using the breath holding protocol and the new IH protocol are shown in Fig. 4.12.

Recall in our previous breath holding protocol used in Chapter 4.4.1, we observed that some participants have their HR and SpO_2 correlated due to the reaction of the cardiac system during breath holding. This is manifested in the histogram shown in the left panel of Fig. 4.13, where 79% (22/28) of the participants' SpO_2 and HR have an absolute

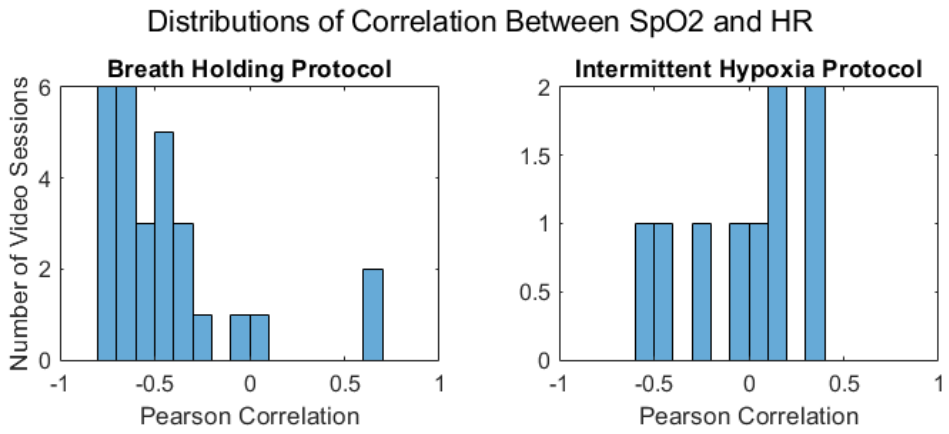


Figure 4.13: Comparison of the correlations between HR and SpO_2 from the breath-holding protocol and the intermittent hypoxia protocol.

correlation greater than a threshold of 0.4. While in the new intermittent hypoxia protocol, as shown in the left panel of Fig. 4.13, only 22% (2/9) have an absolute correlation greater than 0.4. This indicates that this new IH protocol induces less correlation between HR and SpO_2, serving as a new scenario to test the robustness of our proposed algorithm.

SpO_2 Prediction Performance:

The SpO_2 prediction is conducted in the participant-specific manner. The first video session of each participant is used for training and validation, and the second and third video sessions are used for testing. SVR is used for regression. Fig. 4.14 shows the training and testing results for the three participants. For Participant IH-1, the variation in the reference SpO_2 values is small with the lowest SpO_2 being 96% during the video sessions, resulting in no obvious dips in the SpO_2 trend. This may be due to the interpatient variability in tolerance to hypoxia. Thus, even though his predicted test SpO_2 signals do not follow the trend of reference SpO_2 well (with ρ being 0.22 and -0.26, respectively), the MAEs are less than 0.65%. Overall speaking, the test MAEs are within 1.64% while the dips of some of the SpO_2 trends are not captured well, such as Participant IH-2. From

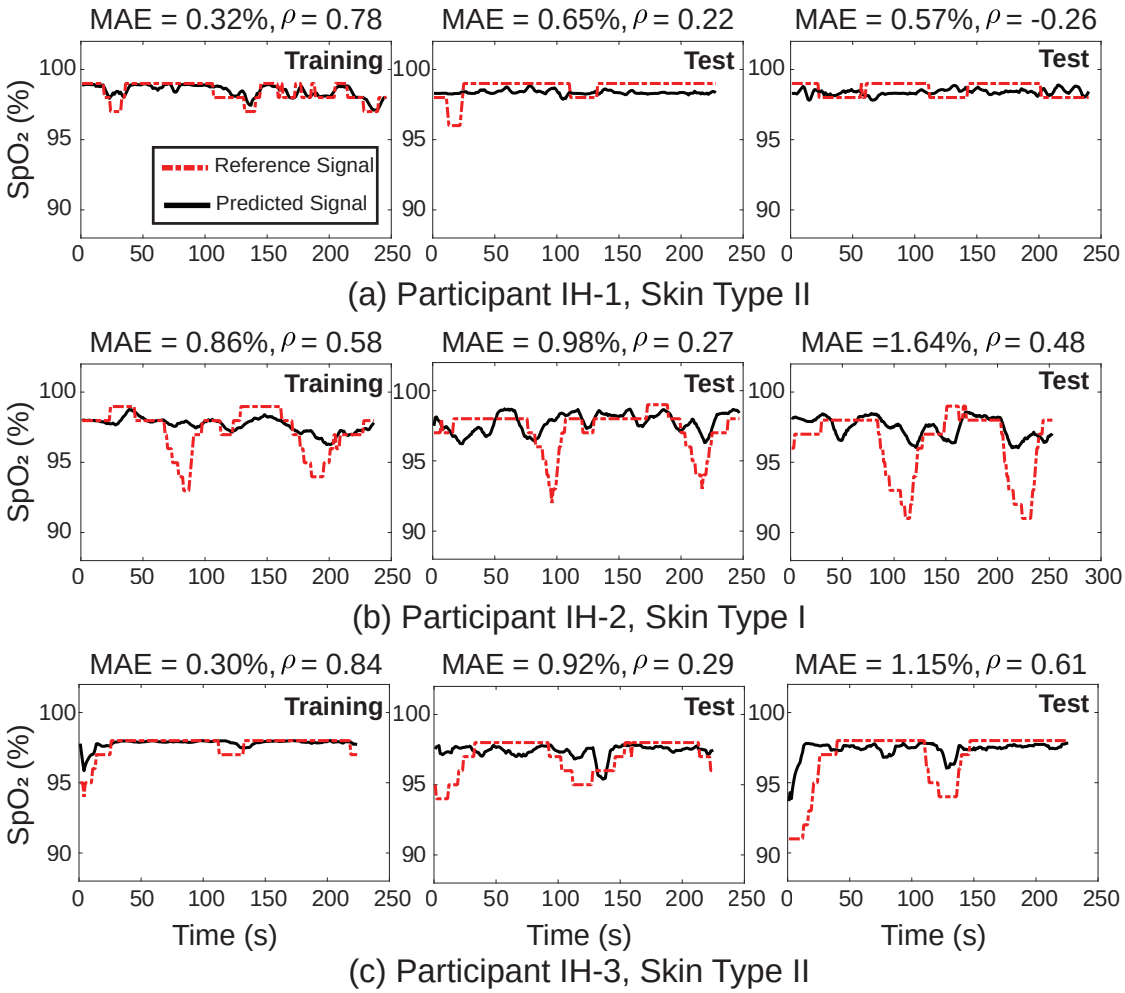


Figure 4.14: Predicted SpO_2 signals using SVR are shown for all participants from the IH protocol. The reference SpO_2 is in red dash lines and the predicted SpO_2 is in solid black lines. The higher the correlation ρ and the lower the MAE, the better the predicted SpO_2 captures the trend of the reference signal.

the limited data that we have collected so far with the IH protocol, our proposed algorithm achieved reasonable results and the results need to be verified and further improved with more data collected.

Discussions for Future Development:

From the comparison of SpO_2 distributions between the breath-holding protocol and the IH protocol shown in Fig. 4.12 and SpO_2 trends in Fig. 4.14 versus those in Fig. 4.5, we observe that the drop of SpO_2 does not get deeper and wider with the new IH protocol as described in the literature with similar IH protocols [43, 62]. The differences between our IH protocol and that in the literature mainly lie in the duration of the hypoxia period (in each episode and overall), its relative duration to the normoxia phase, and the fraction of the inspired oxygen. For example, in [62], the hypoxia environment induced by the FiO_2 (the fraction of inspired oxygen) protocol lasts for consecutively 16 minutes on average per participant to create a much wider range of SpO_2 from 61% to 100%, though the fraction of oxygen is unclear in the paper. In [43], the IH experiment is conducted 5 sessions per week throughout a 3-week duration. They found the most prominent decrease of SpO_2 was 10% on average, which happened in week 3 with 5 times of 5-minute hypoxia provoked by 12% oxygen interspersed with 3-minute normoxia intervals in each session.

With the IH protocols applied in the literature and the suggestions of a proper level of hypoxia and duration that lead to safe and positive effects and therapeutic potential of intermittent hypoxia [97], we consider the following modifications in our future design of protocol with advice and supervision from physicians to prevent adverse effects to the participants:

- having a relatively longer hypoxia period (e.g., increase from 25 seconds to several minutes); and/or
- inducing a modest fraction of inspired oxygen (e.g., 9% to 16%) [97] that can match the increased duration of the hypoxia period.

With the longer and larger decrease in SpO_2 values created by the updated protocol, we may have more meaningful training samples and better take advantage of the IH protocol.

4.6 Chapter Summary

This chapter presents a contact-free method of measuring blood oxygen saturation from hand videos captured by smartphone cameras. The whole algorithm pipeline includes 1) receiving video of the hand of a subject captured by a regular RGB camera of a smartphone; 2) extracting a region of interest of the hand video; 3) performing feature extraction of the region of interest based on spatial and temporal data analysis of more than two color channels; and 4) estimating a blood oxygen saturation level of the subject from the features. The key contributions of this chapter are mainly focused on the proposed feature engineering method, which is a synergistic combination of several key components, including the multi-channel ratio-of-ratios feature set, the narrowband filtering that adaptively centered at heart rates, and the accurately estimated heart rate. We have seen encouraging results of a mean absolute error of 1.26% with a commercial pulse oximeter as the reference, outperforming the conventional ratio-of-ratios method by 25%. We have also analyzed the impact of the sides of the hand and skin tones on the SpO_2 estimation. We have found that, given our collected dataset, the palm side performs well regardless of

the skin tone; for palm-up cases, we do not observe significant performance differences between lighter and darker skin tones.

Chapter 5

Optophysiological Model Guided Neural Networks for Contactless Blood Oxygen Estimation From Hand Videos

5.1 Introduction

Deep learning has demonstrated promising performance in camera-based physiological measurements, such as heart rate, breathing rate, and body temperature [26, 101, 127, 162]. An end-to-end convolutional attention network was proposed in [26] to estimate the blood volume pulse from face videos. Frequency analysis is then conducted on the estimated pulse signal for heart rate and breathing rate tracking. The study in [101] demonstrates that the heart rate can be directly inferred using a convolutional network with spatial-temporal representations of the face videos as its input. Mobile applications have been developed to utilize CNNs to measure body temperature from facial images [162].

Deep learning for SpO_2 monitoring from videos is still in its early stage. Ding *et al.* [37] proposed a convolutional neural network architecture for contact-based SpO_2 monitoring with smartphone cameras. Even though the work in [37] showed better per-

formance than the conventional ratio-of-ratios method, their technique requires the users' fingertips to be in contact with the illuminated flashlight and camera, which not only may lead to a sense of burning for a continuous period of time but also raises sanitation concerns, especially if the sensing device is shared by multiple participants during pandemics.

The video-based contactless methods for physiological signal sensing provide a comfortable and an unobtrusive option of monitoring SpO_2 and have the potential to be adopted in health screening and telehealth. In Chapter 4, we have taken advantage of the contact-free sensing from a regular RGB camera as well as the well-known two-channel RoR mechanism from pulse oximeters for accurate SpO_2 estimation. In particular, we proposed a strategic use of the hand video data by performing spatial and temporal data analysis of more than two color channels. Under the umbrella of the synergistic framework that takes advantage of both biophysical imaging principles and the availability of participants' video and SpO_2 data to learn and determine the details for obtaining SpO_2-relevant features and making SpO_2 estimation, Chapter 4 determined the specific features and the related detailed parameters **explicitly** from the biophysical imaging principles, while in this chapter, we propose to use these principles to guide the design of neural network architectures to "learn" the specific SpO_2-relevant features from the input video signals with a data-driven **implicit** approach and perform SpO_2 estimation in a holistic manner. Compared to the principled signal processing scheme for feature engineering proposed in Chapter 4, the neural network based schemes proposed in this chapter learn features implicitly from data and use synergy with the principled methodology to guide the selection of the neural network architectures.

Specifically, inspired by the optophysiological model for SpO_2 measurement [117, 142, 152], in this chapter, we develop convolutional neural networks (CNN) based SpO_2 estimation schemes designed based on the optophysiological models for a better explanation, wherein data-driven feature extraction and estimation of the blood oxygen saturation level comprise implementing a combination of spatial averaging, color channel mixing, and temporal trend analysis. The schemes analyze the videos of a participant's hand captured by regular RGB cameras in a contactless way, which is convenient and comfortable for users and can protect their privacy and allow for keeping face masks on.

5.2 Proposed Optophysiology-Guided Neural Network Method for estimating SpO_2 From Videos

Fig. 5.1 is an overview of the system design. First, the ROI, including the palm and the back of the hand, is extracted from the smartphone captured videos. Second, the ROI is spatially averaged to produce R, G, and B color time series. Next, the three color-channel signals are fed into an optophysiology-inspired CNN to extract features to achieve more explainable and accurate SpO_2 predictions.

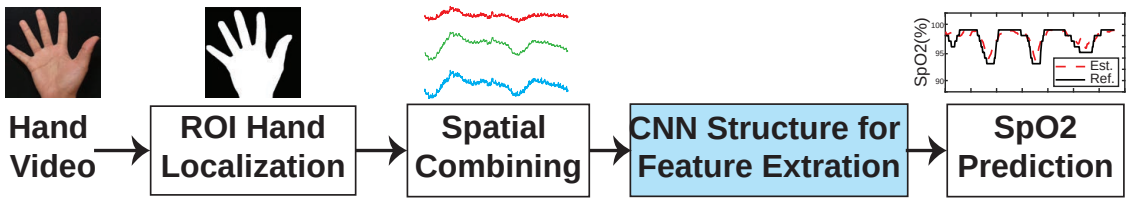


Figure 5.1: Proposed neural network based contactless SpO_2 estimation method. Three color time series are extracted from the skin area of a hand video by spatial averaging and are then fed into an optophysiology-inspired neural network to extract features by color channel mixing and temporal analysis for SpO_2 prediction.

5.2.1 Extraction of Skin Color Signals

The physiological information related to SpO_2 is embedded in the color of the reflected/reemitted light from a person's skin. Hence, a preprocessing step that precisely extracts the color information from the skin area is crucial to the design of an effective SpO_2 estimation method. For each participant's video, we aim to extract the R, G, and B time series and refer to these 1-D time series as *skin color signals*. As Chapter 4.3 explained, the ROI of the skin pixels is separated using Otsu's method [103] which determines a threshold that best separates the skin pixels from the background by minimizing the variance within the skin and non-skin classes in the Cr axis of the YCbCr color space [18]. Once the ROI corresponding to the hand is located, the R, G, and B time series are generated by spatially averaging over the values of skin pixels for each frame of the video.

In this chapter, the skin color signals are split up into 10-second segments using a sliding window with a step size/stride of 0.2 seconds to serve as the inputs for neural networks. From an optophysiological perspective, the reflected/reemitted light from the skin for the duration of one cycle of heartbeat, i.e., 0.5–1 seconds for a heart rate of 60–120 bpm, should contain almost the complete information necessary to estimate the instantaneous SpO_2 [121]. In our system design, we use longer segments to add resilience against sensing noise. Since the segment length is one order of magnitude longer than the minimally required length to contain the SpO_2 information, we can use a fully-connected or convolutional structure to adequately capture the temporal dependencies without resorting to a recurrent neural network structure.

5.2.2 Neural Network Architectures

The previous neural network work for SpO_2 prediction mainly explored prediction, but not the model explainability [37]. Explainability/interpretability is highly desirable in many applications yet often not sufficiently addressed, partly due to the black box nature of neural networks.

From a healthcare standpoint, explainability is a key factor that should be taken into account at the beginning of the design of a system. To extract features from the skin color signals and estimate SpO_2, we propose three physiologically motivated neural network structures. These structures are inspired by domain knowledge-driven physiological sensing methods and designed to be physically explainable. For heart rate sensing [101, 166] and respiratory rate sensing [96, 126], the RGB skin color signals are often combined first to form one "rPPG" signal followed by temporal feature extraction, as is done in the plane-orthogonal-to-skin (POS) algorithm [149]. In contrast, for conventional SpO_2 sensing methods such as the ratio-of-ratios [152], the temporal features are extracted first (i.e., extracting AC and DC from a time window for each color channel) and the color components are combined at the end (i.e., taking the ratio and pairwise ratio of ratios) before doing regression fitting. Our proposed neural network structures explore different arrangements of channel combination and temporal feature extraction. We want to systematically compare the performance of our explainable model structures.

Color Channel Mixing Followed by Temporal Analysis: In Model 1, shown as the leftmost structure depicted in Fig. 5.2, we combine the color channels first using several channel combination layers and then extract temporal features using temporal convolution

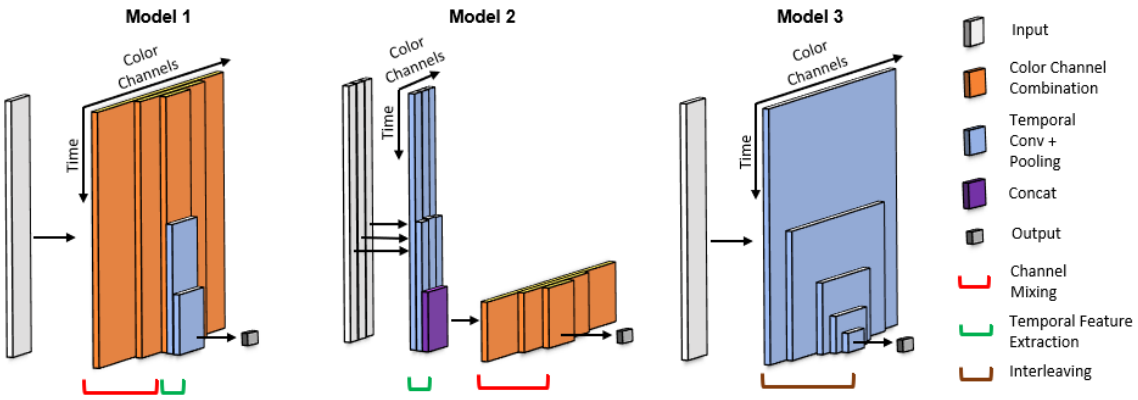


Figure 5.2: Proposed network structures for predicting SpO_2 levels from a fixed-length segment of skin color signals. We highlight the differences among the three model configurations instead of showing the exact model structures. Model 1 combines the RGB channels before temporal feature extraction. Model 2 extracts the temporal features from each channel separately and fuses them toward the end. Model 3 interleaves color channel mixing and temporal feature extraction.

and max pooling. A channel combination layer first linearly combines the C_{in} input channels/vectors into C_{out} activation vectors and then applies a rectified linear unit (ReLU) activation function to obtain the output channels/vectors. Mathematically, the channel combination layer is described as follows:

$$\mathbf{V} = \sigma(\mathbf{W}\mathbf{U} + \mathbf{b}\mathbb{1}^T), \tag{5.1}$$

where $\mathbf{U} \in \mathbb{R}^{C_{\text{in}} \times L}$ is the input comprised of C_{in} time series/vectors of length L. The initial channel combination layer has an input of three channels with 300 points along the time axis. $\mathbf{W} \in \mathbb{R}^{C_{\text{out}} \times C_{\text{in}}}$ is a weight matrix, where each of the C_{out} rows of the matrix is a different linear combination for the input channels. A bias vector $\mathbf{b} \in \mathbb{R}^{C_{\text{out}}}$ contains the bias terms for each of the C_{out} output channels, which ensures that each data point in the created segment of length L has the same intercept. $\mathbb{1}^T \in \mathbb{R}^{1 \times L}$ is a row vector of all ones. The nonlinear ReLU function $\sigma(x) = \max(0, x)$ is applied elementwise to

the activation map/matrix. The output of the channel combination layer $\mathbf{V} \in \mathbb{R}^{C_{\text{out}} \times L}$ contains C_{out} channels of nonlinearly combined input channels.

The channel mixing section concatenates multiple channel combination layers with decreasing channel counts to provide significant nonlinearity. The output of the last channel combination layer has seven channels. After the channel mixing, for temporal feature extraction, we utilize multiple convolutional and max pooling layers with a downsampling factor of two to extract the temporal features of the channel-mixed signals. When there are multiple filters in the convolutional layer, there will also be some additional channel combining with each filter outputting a channel-mixed signal. Finally, a single node is used to represent the predicted SpO_2 level. This model has three channel combination layers, three feature extraction layers, and a total of 34K trainable parameters.

Temporal Analysis Followed by Color Channel Mixing: In Model 2, which is the middle structure depicted in Fig. 5.2, we reverse the order of color channel mixing and temporal feature extraction from that in Model 1. The three color channels are separately fed for temporal feature extraction. The convolutional layers learn different features unique to each channel. At the output of the temporal feature extraction section, each color channel has been downsampled to retain only the important temporal information.

The color channels are then mixed together in the same way as described for Model 1 before outputting the SpO_2 value. This model has three channel combination layers, 2 feature extraction layers, and a total of 12K parameters.

Interleaving Feature Extraction and Channel Mixing: In our third model, we explore the possibility of interleaving the color channel mixing and temporal feature extraction steps. As illustrated by the rightmost structure depicted in Fig. 5.2, the input is first put

through a convolutional layer with many filters and then passed to max pooling layers, resulting in feature extraction along the time as well channel combinations through each filter. The number of filters is reduced with each successive convolutional layer, gradually decreasing the number of combined channels and downsampling the signal in the time domain. This model has 4 layers and a total of 307K parameters.

Loss Function and Parameter Tuning. We use the root-mean-squared-error (RMSE) as the loss function for all models. During training, we save the model instance at the epoch that has the lowest validation loss. The neural network inputs are scaled to have zero mean and unit variance to improve the numerical stability of the learning. The parameters and hyperparameters of each model structure were tuned using the HyperBand algorithm [86], which allows for a faster and more efficient search over a large parameter space than grid search or random search. It does this by running random parameter configurations on a specific schedule of iterations per configuration and uses earlier results to select candidates for longer runs. The parameters that are tuned include the learning rate, the number of filters and kernel size for convolutional layers, the number of nodes, the dropout probability, and whether to do batch normalization after each convolutional layer.

5.3 Experimental Results

5.3.1 Dataset and Capturing Conditions

Our proposed models are evaluated on a self-collected dataset that is studied in Chapter 4. To recapitulate, the dataset consisted of two sessions of hand video record-

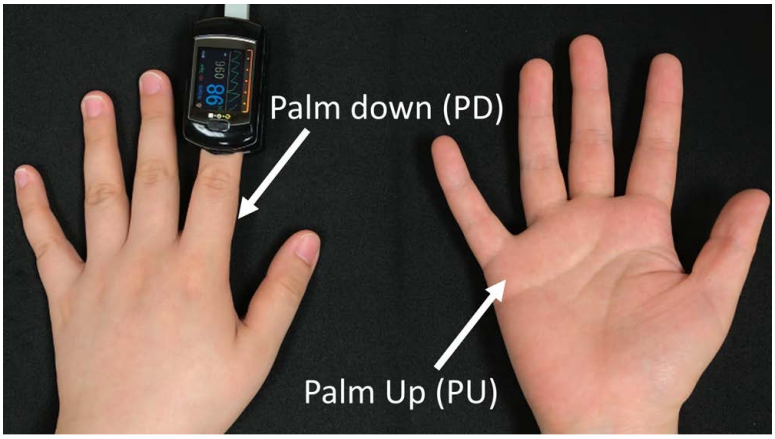
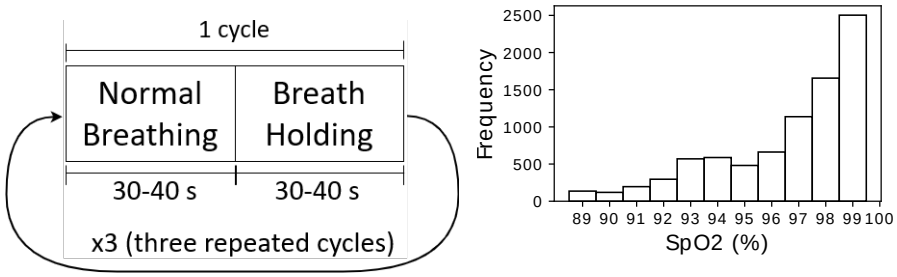


Figure 5.3: Illustration of two hand-video capturing positions. Left hand: palm down (PD). Right hand: palm up (PU).

ings and simultaneously recorded reference SpO_2 data from each of the 14 participants, of which there were six males and eight females between the ages of 21 and 30. The distribution of the participants' skin types is as follows: Two participants of type II, eight participants of type III, one participant of type IV, and three participants of type V. This research was using protocol #1376735 approved by the University of Maryland Institutional Review Board (IRB).

Each participant was asked to place his/her hands still on a table to avoid hand motion. Their palm of the right hand and the back of the left hand are facing the camera, as illustrated in Fig. 5.3. We refer to these two hand-video capturing positions as *palm up (PU)* and *palm down (PD)*, respectively. Each participant was asked to follow the breathing protocol outlined in Fig. 5.4(a). The participant breathes normally for 30–40 seconds, exhales all the way, and then holds his/her breath for 30–40 seconds. This process is repeated three times for each session. The collected SpO_2 value distribution is shown in Fig. 5.4(b).

In this chapter, we increase the data size by interpolating the reference SpO_2 signal



(a) (b)

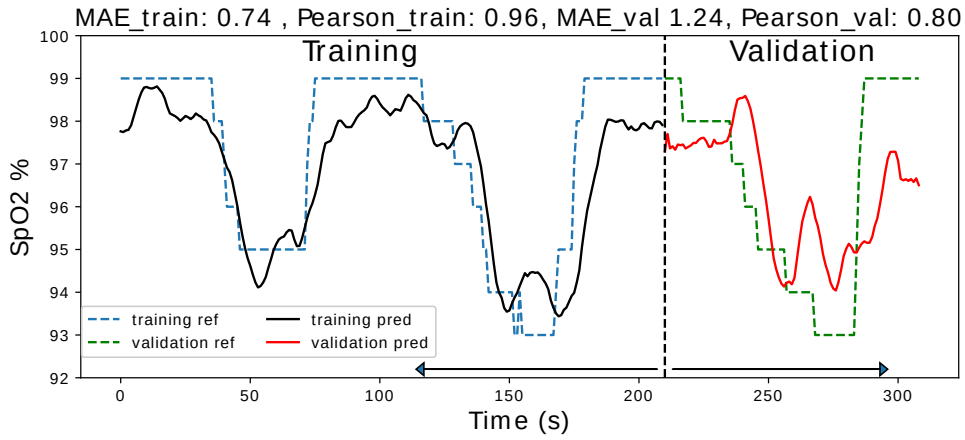
Figure 5.4: (a) Breathing protocol that participants were asked to follow, including 3 cycles of normal breathing and breath holding. (b) Histogram of SpO_2 values in the collected dataset.

to 5 sample points per second to match the segment sampling rate (Chapter 5.2.1) using a smooth spline approximation [50]. Each RGB segment and SpO_2 value pair is fed into our models as a single data point, the models output a single SpO_2 estimate per segment. To evaluate a model on a video recording, the model is sequentially fed with all RGB segments from the recording to generate a time series of preliminarily predicted SpO_2 values. All predictions greater than 100% SpO_2 are clipped to 100% based on physiological knowledge. A 10-second long moving average filter is applied to generate a refined time series of predicted SpO_2 values.

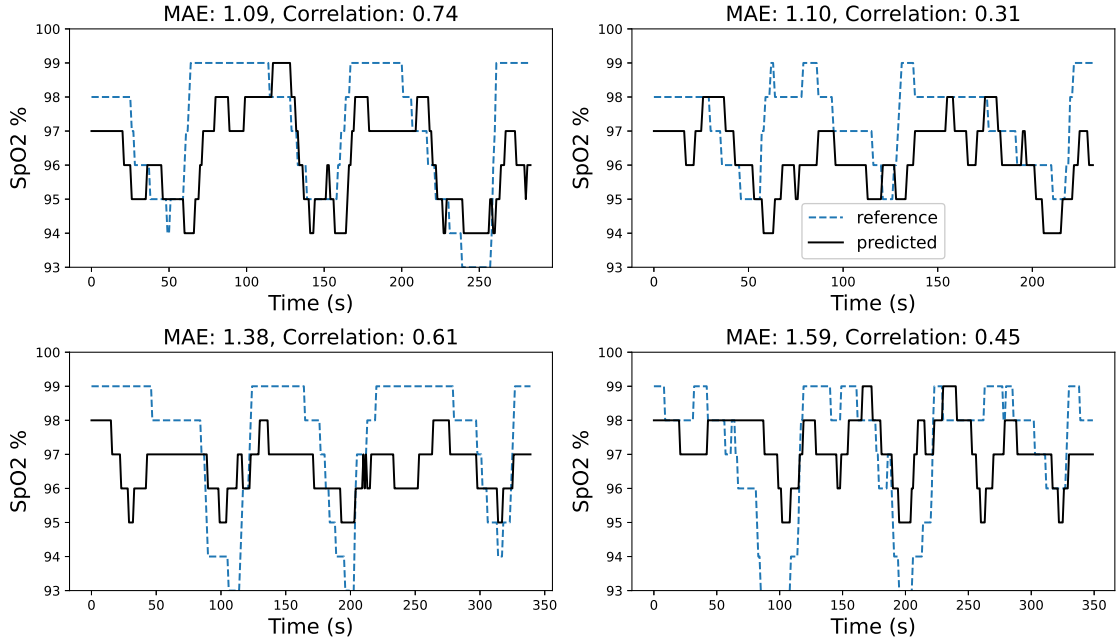
5.3.2 Participant-Specific Results

To investigate how well the proposed models could learn to estimate a specific individual's SpO_2 from his/her own data, we first conducted participant-specific experiments, that is, we learn individualized models for each participant.

Experimental Setting: Two recordings per participant were captured with at least 15 minutes in between. One recording is used for training and validation of the model and



(a)



(b)

Figure 5.5: (a) Training vs. validation predictions. (b) Test predictions of varying performance with reference SpO_2. The higher the Pearson correlation, the better the prediction captures the reference SpO_2 trend. The lower the MAE, the better the prediction captures the dips in SpO_2.

the remaining recording is for testing. An example of the training and validation predictions curves is shown in Fig. 5.5(a). Each recording contains three breathing cycles, for each training/validation recording, the first two breathing cycles are taken for training and the third cycle is used for validation. Splitting the recordings into cycles instead of

randomly sampling the 10-sec overlapping RGB segments ensures that there are no overlapping segments of data between the training and validation set. Example test prediction curves and their correlation and mean-absolute-error (MAE) are shown for reference in Fig. 5.5(b). It should be noted that if the correlation is low, e.g., a constant temporal estimate, then the MAE and RMSE metrics are less meaningful. For the participant-specific experiments, due to the small dataset size, we augment the training and validation data by sampling with replacement. This is an example of the bootstrapping data reuse strategy [67, Chapter 5]. The oversampling also helps address the imbalance in SpO_2 data values that is shown in Fig. 5.4(b).

In each experiment, the model structure and hyperparameters are first tuned using the training and validation data. Once the model has been tuned, we train multiple instances of the model using the best tuned hyperparameters. Between each instance, we vary the random seed used for model weights initialization and random oversampling. Each model instance is evaluated on the training/validation recording, and the model instance that achieves the highest validation RMSE is selected for evaluation on the test recording. This model is then evaluated on the test recording to obtain the final test results.

Results: Table 5.1 shows the performance comparison of our proposed models with the prior-art model from Ding *et al.* [37]. To the best of our knowledge, Ding *et al.*'s model is the only convolutional neural network structure that has been tried for contact-based SpO_2 estimation. Its structure is similar to our Model 3 but with fewer layers. We also compare with the classic ratio-of-ratios method proposed by Scully *et al.* [117]. The performance is measured in Pearson's Correlation, mean absolute error (MAE), and root mean square

	Hand Mode	Correlation		MAE (%)		RMSE (%)	
		Median	IQR	Median	IQR	Median	IQR
Model 1 (Proposed)	PD	0.41	0.40	2.12	0.91	2.51	0.78
	PU	0.39	0.37	2.16	1.80	2.70	2.09
Model 2 (Proposed)	PD	**0.46**	0.44	2.09	1.32	2.52	1.63
	PU	**0.41**	0.32	1.96	0.68	2.48	0.89
Model 3 (Proposed)	PD	0.44	0.40	**1.93**	1.11	2.48	1.31
	PU	**0.41**	0.46	**1.81**	1.83	2.43	2.44
Scully *et al.* [117]	PD	0.08	0.37	1.94	0.92	**2.22**	0.77
	PU	0.19	0.24	2.01	0.80	**2.36**	0.78
Ding *et al.* [37]	PD	0.38	0.39	3.25	2.85	3.83	3.24
	PU	0.34	0.56	3.40	3.16	4.58	3.12

Table 5.1: Performance comparison of each model structure for participant-specific experiments. Results are given as the test median and IQR of all participants.

error (RMSE), and the results of each condition are summarized in the median and interquartile range (IQR). IQR quantifies the spread of an empirical distribution of a set of data points by computing the difference between the first quartile and the third quartile of the distribution.

Table 5.1 reveals that Model 2 achieves the best correlation in both PD and PU cases, whereas Model 3 achieves the best MAE and a comparable correlation with Model 2, suggesting that Model 2 and Model 3 are comparably the best in the individualized learning. Even though the method proposed in Scully *et al.* [117] achieves the best (lowest) RMSE, its correlations are the worst (lowest). This suggests that the classic ratio-of-ratios method cannot track the trend of SpO_2 well using the contactless measurement by smartphone. All of our model configurations outperform Ding *et al.* [37]. For example, in the PU case for Model 3, the correlation is improved from 0.34 to 0.41 and the MAE is lowered from 3.40% to 1.81%. It is worth noting that the international standard for clinically acceptable pulse oximeters allows an error of 4% [65], and our estimation errors are all within this range.

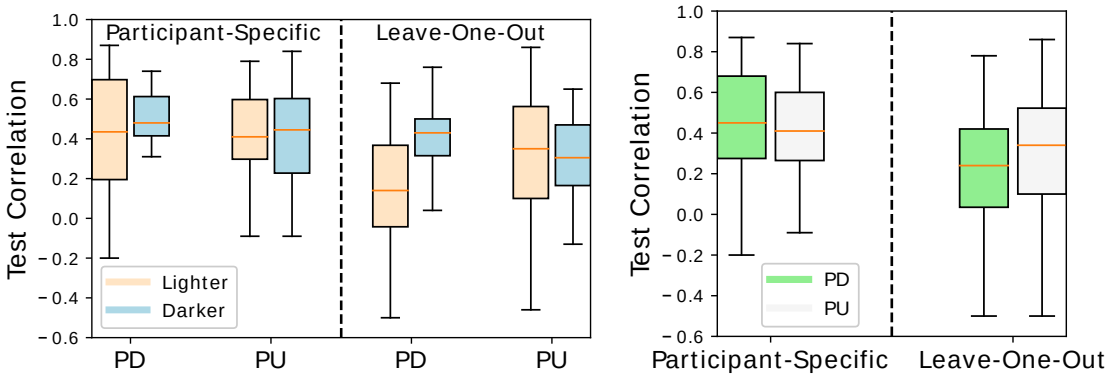


Figure 5.6: Boxplots comparing distributions of correlations for (a) lighter vs. darker skin types, and (b) PD vs. PU for all skin types. The PD results are better for darker skin tones in both the participant-specific and leave-one-out cases.

There are two factors, including the skin type and the side of the hand, which might influence the performance of SpO_2 estimation. We therefore investigate the following two questions: (1) Whether the different skin types matter in PU or PD cases, and (2) whether the side of hand matters in lighter skin (types II + III) or darker skin (types IV + V). The box plots in Fig. 5.6 summarize the distributions of the test correlations from all the three proposed models in PU and PD modes of (a) lighter-skin and darker-skin participants, and (b) all participants.

Bayesian statistical test: We use Bayesian statistical tests to further analyze the results in Fig. 5.6 by providing a probabilistic assessment of whether the results from two groups being compared have the same mean [78–82]. We avoid using the popular t-test because it makes only a binary decision due to its lack of direct information about the probability of difference between group means of the given data [78]. In contrast, the Bayesian statistical test computes the posterior distribution of difference between the two group means to quantify its certainty of possible values [82]. The decision rule of the Bayesian

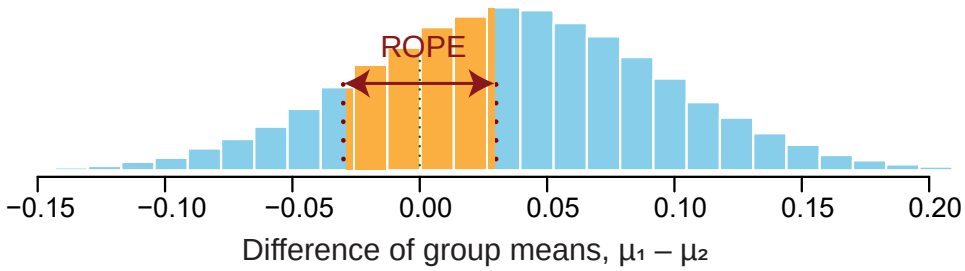


Figure 5.7: Posterior distribution of the difference of group means. This shows an example of an undecided case of the Bayesian statistical test given that the ROPE of zero difference is set to $[-0.03, 0.03]$ and 33% of the posterior distribution falls within the ROPE. The percentage of coverage can be used to quantify the certainty that two groups have the same mean.

statistical test for the null hypothesis that the two groups have the same mean can be stated as follows given the *region of practical equivalence (ROPE)* of zero difference [80]:

- (*Accepted*): If the percentage of the posterior distribution of the group-mean difference inside the ROPE is sufficiently high (e.g., greater than 95% [80]), then the null hypothesis is accepted.

- (*Rejected*): If the percentage of the posterior distribution of the group-mean difference inside the ROPE is sufficiently low (e.g., less than 2.5% [80]), then the null hypothesis is rejected.

- (*Undecided*): When the null hypothesis is neither accepted nor rejected, the percentage of the posterior distribution of the group-mean difference inside the ROPE can be used to quantify the certainty that two group means are the same. One example is shown in Fig. 5.7.

To conduct the Bayesian statistical tests, we use an R statistical package named BEST [110]. To determine the ROPE on the difference between the means, we use Cohen's established

convention that the ROPE of small standardized mean difference is [-0.1,0.1] [79, 81]. Given the standard deviation being 0.3 of our data, the ROPE for the difference of means of our data is scaled to [-0.03, 0.03].

To answer question (1) about the impact of the skin type on the prediction performance, we focus on the left panel of Fig. 5.6(a). For the PD case, only 14% of the posterior distribution of the difference between the means of the lighter and darker skin groups falls in the ROPE. For the PU case, 23% of the posterior distribution falls in the ROPE. This suggests that it is highly credible to conclude that the skin type makes a difference in SpO_2 prediction, and the difference is more certain to be observed when using the back of the hand as the ROI compared to using the palm.

To answer question (2), we first focus on the left panel of Fig. 5.6(b) when participants of all skin colors are considered together. 33% of the posterior distribution of the difference of means between PU and PD cases falls in the ROPE. We then zoom into the darker skin group as shown in the left panel of Fig. 5.6(a), only 15% of the posterior distribution of the difference of means between PD and PU cases falls in the ROPE, whereas for the lighter skin group, 31% of the posterior distribution falls in the ROPE. This implies that it is highly credible that the side of the hand may have some impact on SpO_2 prediction, especially when concerning mainly the darker skin group.

5.3.3 Leave-One-Participant-Out Results

To investigate whether the features learned by the model from other participants are generalizable to new participants whom it has not seen before, we conduct leave-one-

participant-out experiments. For each experiment, when testing on a certain participant, we use all the other participant's data for training and leave the test participant's data out. The recordings from all the non-test participants are used for participant-wise cross-validation to select the best model structure and hyperparameters. The selected model is evaluated on the two recordings of the test participant, whose data was never seen by the model during training.

Table 5.2 shows the performance comparison of each model in leave-one-participant-out experiments. Model 1 achieved the best performance in terms of correlation and achieved the best MAE and RMSE for the PU case. Similar to the participant-specific case, the classic ratio-of-ratios method proposed in Scully *et al.* [117] achieved better MAE and RMSE results for the PD case but the correlation result was low, suggesting that the model achieved low error by simply predicting a nearly constant SpO_2 near the middle of the SpO_2 range. The best performance of Model 1 in the leave-one-participant-out experiment may imply that the features extracted after combining the color channels at the beginning of the pipeline can be generalized better to unseen participants than the features extracted before channel combination or through interleaving as in Models 2 or 3.

In the participant-specific case, the model is specifically tailored to the test individual, whereas the leave-one-participant-out case is more difficult because the model needs to accommodate for the variation in the population. As expected, in Fig. 5.6, we observe that the overall results from the leave-one-participant-out experiments do not match those from the participant-specific experiments. Because of the modest size of the dataset, the model has not seen as diverse data as a larger and richer dataset would offer. The generalization capability to new participants can be improved when more data is available.

	Hand Mode	Correlation		MAE (%)		RMSE (%)	
		Median	IQR	Median	IQR	Median	IQR
Model 1 (Proposed)	PD	**0.33**	0.42	2.33	1.07	3.07	1.52
	PU	**0.46**	0.36	**1.97**	0.80	**2.32**	0.87
Model 2 (Proposed)	PD	0.15	0.50	2.43	0.94	3.35	1.11
	PU	0.33	0.39	2.08	0.73	2.41	0.71
Model 3 (Proposed)	PD	0.23	0.38	2.48	1.18	2.98	1.33
	PU	0.27	0.31	2.02	1.03	2.54	1.28
Scully *et al.* [117]	PD	0.05	0.43	**2.08**	0.65	**2.44**	1.14
	PU	0.01	0.54	2.08	0.60	2.43	1.20
Ding *et al.* [37]	PD	0.11	0.56	3.19	1.61	3.76	1.52
	PU	0.26	0.42	2.43	1.22	2.85	1.51

Table 5.2: Performance comparison of each model structure in leave-one-participant-out experiments. Results are given as the test median and IQR of all participants.

We now revisit the two research questions raised in Section 5.3.2 under the leave-one-participant-out scenario. First, we analyze the impact of skin type given the same side of the hand. From the right panel of Fig. 5.6(a), in the PD case, only 0.04% of the posterior distribution of the difference of means between lighter and darker skin groups is within the ROPE, suggesting that the null hypothesis is rejected and the darker skin group outperforms the lighter skin group. In the PU case, 18% of the posterior distribution falls within the ROPE. This observation is consistent with the participant-specific experiments that when using the back of the hand as the ROI, the skin color is more credible to be a factor in the accuracy of SpO_2 estimation than using the palm.

Second, we analyze the impact of the side of the hand for two skin color groups. For the darker skin group shown in the right panel of Fig. 5.6(a), only 9% of the posterior distribution of the difference of means of the PU and PD cases falls in the ROPE. This shows that there is high uncertainty in the estimate of zero difference, which is consistent with the results from the participant-specific experiments. However, unlike the participant-specific experiments, for the lighter skin group, 0.2% of the posterior distribu-

tion of the difference of means between PU and PD cases falls in the ROPE. This suggests that the null hypothesis is rejected and that the PU outperforms the PD in the lighter skin group. As for the mixed group illustrated in the right panel of Fig. 5.6(b), only 8% of the posterior distribution of the difference of means falls in the ROPE, suggesting that there is a high uncertainty to conclude that PU and PD cases are comparable.

This different generalization capability in the PU and PD cases may be attributed to the skin color difference between the palm and the back of the hand. The color of the back of the hand tends to be darker than the color of the palms and has larger color variation among participants due to different degrees of sunlight exposure. In contrast, the color variation of the palms is much milder among participants. Furthermore, in the participant-specific experiments, the individualized models learn the traits of the skin type and the side of the hand from each participant, whereas, in the leave-one-participant-out experiments, the learned model must capture the general characteristics of the population.

5.3.4 Ablation Studies

To justify the use of nonlinear channel combinations and convolutional layers for temporal feature extraction in our proposed models, we conduct two ablation studies comparing the performance of these model components to other generic ones. We focus on the PU case to avoid the uncontrolled impact of such factors as skin tone and hair. In the first ablation study, we compare nonlinear to linear channel combinations. We create a variant of Model 1 with only a single linear channel combination layer with no activation function and repeat the leave-one-participant-out experiments. In the second study, we

Method		ρ	MAE(%)	RMSE(%)
Linear Ch. Comb.	Median	0.46	2.14	2.66
+ Conv. layer for Feat. Extra.	IQR	0.38	0.73	0.93
Nonlinear Ch. Comb.	Median	0.41	2.29	2.66
+ Fully Connec. layer for Feat. Extra.	IQR	0.39	0.63	0.70
Model 1 (Proposed): Nonlinear Ch. Comb.	Median	**0.46**	**1.97**	**2.32**
+ Conv. layer for Feat. Extra.	IQR	0.36	0.80	0.87

Table 5.3: Numerical results of the ablation studies for Model 1 (M1) in the leave-one-participant-out mode. Comparisons among the proposed (nonlinear) M1, modified M1 with only linear channel combinations, and modified M1 with fully connected dense layers instead of convolutional layers are listed. Ablation studies confirm that the nonlinear channel combinations and convolutional layers improve model performance.

compare the performance of using convolutional layers for temporal feature extraction to using fully-connected dense layers. We create this second variant of Model 1 and repeat leave-one-participant-out experiments.

Table 5.3 presents the medians and IQRs specified for numerical comparison of the ablation study. First, we compare the first and the third rows in Table 5.3 for ablation study 1. Our proposed Model 1 achieves a better correlation with a median of 0.46 and IQR of 0.36 and a better RMSE with a median of 2.32 and IQR of 0.87 than its linear channel combination variant. Besides, Model 1 achieves a comparable MAE with a better median of 1.97 but a wider IQR of 0.80. The overall better performance of Model 1 suggests the necessity of using the nonlinear channel combination method. Second, in ablation study 2, we compare the second and the third rows in Table 5.3. We observe that Model 1 outperforms its second variant with fully connected layers for feature extraction with better medians in terms of correlation (0.46 vs. 0.41), MAE (1.97 vs. 2.29), and RMSE (2.32 vs. 2.66), and narrower IQR of correlation. This suggests that convolutional layers are better than fully connected layers for temporal feature extraction.

5.4 Discussions

5.4.1 Contact-based Dataset Testing

We also test our models on the publicly available dataset gathered by Nemcova *et al.* for their SpO_2 estimation work [98]. This dataset consists of contact-based smartphone video recordings where a participant placed a finger on the smartphone camera and was illuminated by the camera flashlight. Participants were asked to breathe normally without following any sophisticated breathing protocol. Each recording lasts about 10 to 20 seconds. The subject for each recording is not identified, so subject-specific and leave-one-participant-out experiments cannot be conducted. There is a single reference SpO_2 value associated with each recording. We used 14 recordings for training and seven recordings for testing and compared them with the modified ratio-of-ratios method proposed in their paper.

As shown in Table 5.4, Models 1 and 2 outperform the method used by Nemcova *et al.* on both the training and test recordings. Model 3 is not able to generalize well from the training set to the test set, which may be due to the small size of the dataset. It should be noted that because the participants were not asked to follow any sophisticated breathing protocol, the dynamic range of SpO_2 values is narrow. These results show that our CNN Models 1 and 2 work well for contact-based video recordings in addition to contactless video recordings.

	MAE (%)		RMSE (%)	
	Training	Test	Training	Test
Model 1	0.86	**1.19**	0.94	**1.36**
Model 2	0.50	1.28	0.59	1.64
Model 3	0.75	3.28	0.99	3.69
Nemcova *et al.* [98]	2.05	2.18	2.24	2.36

Table 5.4: Experimental results of proposed methods on the contact-based video SpO_2 dataset from Nemcova *et al.* [98]. One SpO_2 estimate was output per recording and MAE and RMSE were calculated across all recordings. Models 1 and 2 outperform the method proposed by Nemcova *et al.*, Model 3 was unable to generalize well to the test set.

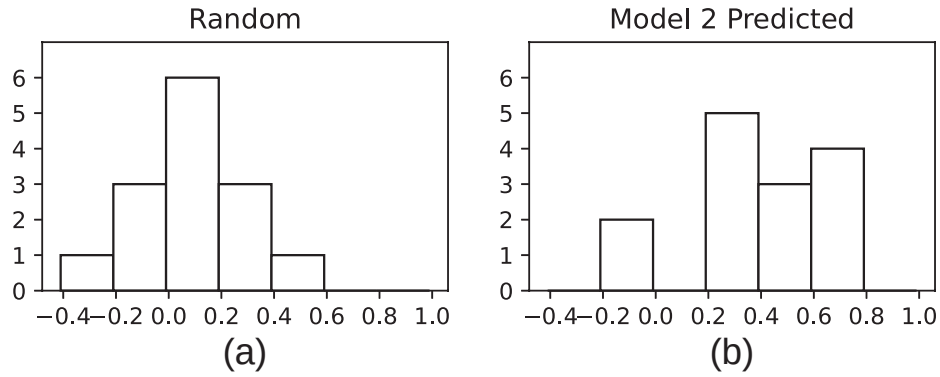


Figure 5.8: Histograms of correlation values between reference SpO_2 signals and (a) randomly generated SpO_2 signals, or (b) SpO_2 signals predicted by neural network Model 2. The correlation distribution for Model 2 is centered much higher than the random guess, confirming Model 2's capability to track SpO_2.

5.4.2 Ability to Track SpO_2 Change

By employing the standard machine learning methodology of training-validation-test split in Section 5.3 to learn neural networks that perform well on unseen data, we have already ensured the generalizability of our models [122, Chapter 11]. As further evidence that our models are capable of outputting meaningful predictions, we compare SpO_2 predictions from our learned models to randomly generated SpO_2 values. For each reference signal, a random prediction signal was generated by choosing SpO_2 values between the minimum and maximum values from the reference signal and applying a moving aver-

age window in the same way as is applied to the neural network predictions. Fig. 5.8(a) shows a histogram of the correlations between the reference SpO_2 signals and the randomly generated predictions and Fig. 5.8(b) shows a histogram of correlations between the reference SpO_2 signals and the predictions generated by Model 2. It is revealed that the neural network with a median correlation of 0.41[1] outperforms random guessing with a median correlation of -0.02, confirming Model 2's capability to track SpO_2.

5.4.3 Visualizations of RGB Combination Weights

To understand and explain what our physiologically inspired models have learned, we conduct a separate investigation to visualize the learned weights for the RGB channels. Our goal is to understand the best way to combine the RGB channels for SpO_2 prediction. Having an explainable model is important for a physiological prediction task like this. Our neural network models can be considered as nonlinear approximations of the hypothetically true function that can extract the physiological features related to SpO_2 buried in the RGB videos. The ratio-of-ratios method, for example, is another such extractor that combines the information from the different color channels at the end of the pipeline. For this experiment, we use the modified version of Model 1 from the ablation studies that has only a single linear channel combination at the beginning. Seeing that using a single linear channel combination did not significantly reduce model performance in the ablation studies, and understanding that the linear component may dominate the Taylor expansion of a nonlinear function, we use only linear combinations for this model

[1] It has been shown in other applications that even low correlation coefficients can be meaningful. For example, in photo response non-uniformity (PRNU) work, the device used to take a photo can be predicted with correlation values below 0.1 [11].

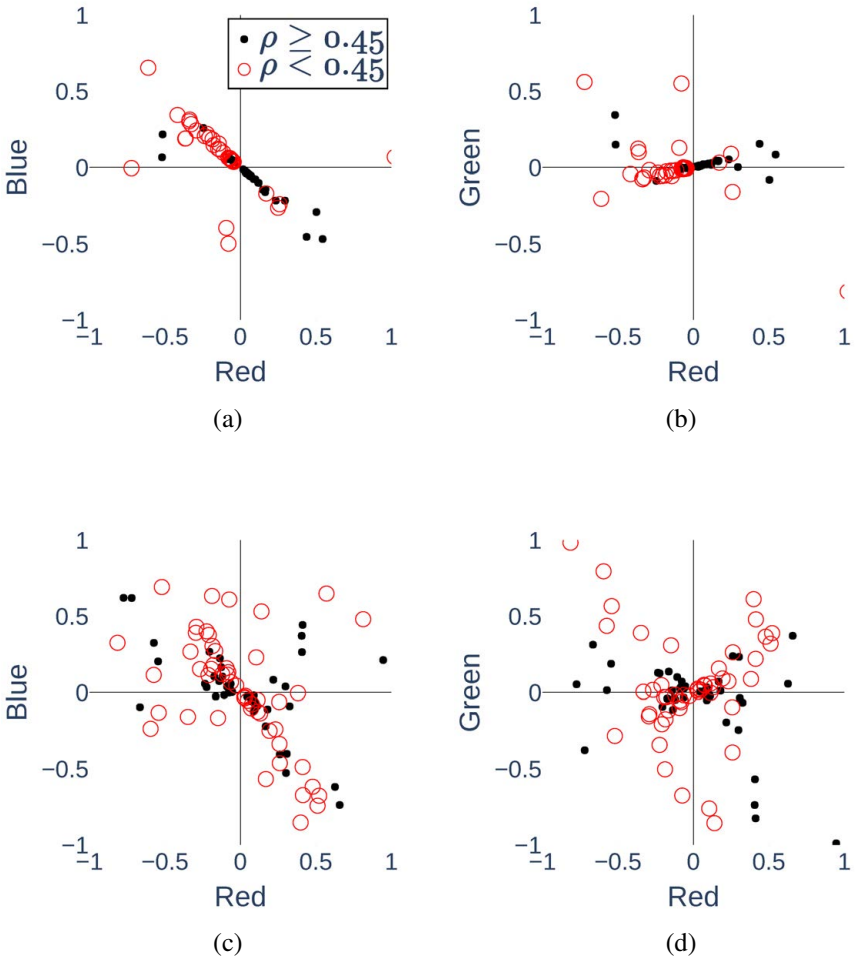


Figure 5.9: Learned RGB channel weights. Plots (a) and (b) are the channel weights learned by different model instances trained on the data of all study participants together, projected onto the RB and RG planes in the RGB space. Plots (c) and (d) are the RB and RG projections of the learned channel weights for model instances trained on random subsets of the participants' data. Each point is color-coded according to the correlation ρ achieved by the instance.

to facilitate more interpretable visualizations.

We have trained 100 different instances of the model on the first two cycles from all the recordings and tested on the third cycle from all recordings. The difference between each instance is that the weights are randomly initialized. The weights for each channel learned by the model instances were visualized as points representing the heads of the linear combination vector in RGB space. Each point is colored according to the average test correlation achieved by the model instance. Figs. 5.9(a) and 5.9(b) show the projections of these points onto the RB and RG planes. The subfigures reveal that the majority of the channel weights lay along certain lines in the RGB space. For the weights on the line, the ratio of the blue channel weight to the red channel weight is 0.87, and the ratio of the green channel weight to the red channel weight is 0.18. It is clear that the red and blue channels are the dominating factors for SpO_2 prediction.

To further verify this result, we repeat this experiment under the following setup: instead of using the data from all participants, for each model instance, we randomly select seven participants and use their data for training and testing. In this case, the difference between each model instance is not only the initialized weights but also the random subset of participants that the model was trained on. Fig. 5.9(d) reveals that most of the better-performing instances (with $\rho \geq 0.45$) have little contribution from the green channel. In Fig. 5.9(c), we again see that most of the points lay on a line in the RB plane, the ratio of the blue channel weight to the red channel weight for these points is 0.80.

These results are in accordance with the biophysical understanding of how light is absorbed by hemoglobin in the blood. Recall that Fig. 1.5 reveals a large difference between the extinction coefficients, or the amount of light absorbed, by deoxygenated and

oxygenated hemoglobin at the red wavelength. There is a significantly smaller difference at the blue wavelength and almost no difference at green. The amount of light absorbed influences the amount of light reflected which can be measured through the camera. A larger difference in extinction coefficients makes it easier to measure the ratio of light absorbed by oxygenated vs. deoxygenated hemoglobin over time. This ratio indicates the level of blood oxygen saturation. Therefore, from a physiological perspective, it makes sense for the neural networks to give larger weight to the red and then blue channels and give little to the green channel. These visualizations indicate that the models are learning physically meaningful features.

5.5 Chapter Summary

We have proposed the first CNN-based work to solve the challenging problem of video-based remote SpO_2 estimation. We have designed three optophysiologically inspired neural network architectures. In both participant-specific and leave-one-participant-out experiments, our models are able to achieve better results than the state-of-the-art method. We have also analyzed the effect of skin color and the side of the hand on SpO_2 estimation and have found that in the leave-one-participant-out experiments, the side of the hand plays an important role, with better SpO_2 estimation results achieved in the palm-up case for the lighter-skin group. We have also shown the explainability of our designed architectures by visualizing the weights for the RGB channel combinations learned by the neural network, and have confirmed that the choice of the color band learned by the neural network is consistent with the established optophysiological methods.

Chapter 6

Conclusions and Future Perspectives

Periodic blood volume change underneath a person's skin induces subtle color variations in the skin area. These subtle changes can be captured by a noninvasive and low-cost optical technique called photoplethysmography (PPG). The methods of PPG measurement have evolved in the past decades from using contact-based devices (e.g., fingertip PPG from the pulse oximeters) to using contactless devices (e.g., remote PPG from the RGB cameras). In this dissertation, we studied the modeling of contact-based and contact-free PPG signals to facilitate its promising applications in cardiovascular signal and vital sign sensing and learning for digital smart health.

In the first part of the dissertation (Ch. 2), we explored the potential of user-friendly and continuous electrocardiogram (ECG) monitoring with the help of contact-based PPG sensors. ECG is a clinical gold standard for non-invasive cardiac monitoring. Given that continuous ECG monitoring in consumer products is challenging, PPG provides a low-cost alternative, though it provides less clinical knowledge compared to ECG. How to leverage the advantages of these two measurement modalities for better and easier healthcare? We approached this problem by first studying the physiological and signal

relationship between PPG and ECG signals, and then inferring the waveform of ECG via the PPG signals based on their relationship. To address this cardiovascular inverse problem, joint dictionary learning frameworks were proposed to learn the mapping that relates the sparse domain coefficients of each PPG cycle to those of the corresponding ECG cycle. This line of research has the potential to fully utilize the easy measurability of PPG and the rich clinical knowledge of ECG for better preventive healthcare.

In the second part of the dissertation (Ch. 3), we developed a physiological digital twin for personalized continuous cardiac monitoring. Digital twins are emerging as a promising framework for realizing precision health for their ability to represent an individual's health status. Using our proposed dictionary learning based algorithm in Ch. 2 as the backbone model, this chapter of the dissertation focused on the problem of inferring ECG signals from PPG signals for continuous precision cardiac monitoring under realistic conditions in which available ECG data is scarce. By performing transfer learning, a generic digital twin model learned from a large portion of paired ECG and PPG data was fine-tuned to precisely infer the ECG from the PPG of a target participant whose available ECG data are scarce. Experimental results showed that the proposed transfer learning method yielded the best ECG reconstruction accuracy compared to other baseline comparison models, which suggested that it could be used as a reliable digital twin for precision continuous cardiac monitoring. In parallel, convolutional neural network based backbone model designs were also proposed based on the underlying physiological process of ECG generation for better explainability.

In the third part of the dissertation (Ch. 4 and Ch. 5), we presented a noncontact method of blood oxygen saturation (SpO_2) monitoring from remote PPG signals captured

by smartphone cameras. SpO_2 is an important indicator of pulmonary and respiratory functionalities. Recent works have investigated how ubiquitous smartphone cameras can be used to infer SpO_2. Most of these works are contact-based, requiring users to cover a phone's camera and its nearby light source with a finger to capture reemitted light from the illuminated tissue. Contact-based methods may lead to skin irritation and cross contamination, especially during a pandemic. Thus, we aimed for contactless methods for SpO_2 monitoring using hand videos acquired by regular RGB cameras of smartphones. Both principled signal processing based method and data-driven neural network based method were proposed for SpO_2 estimation by either explicitly or implicitly extracting features from multi-channel skin color signals with color channel mixing and temporal analysis. Experimental results showed that our proposed methods could achieve better accuracy of blood oxygen estimates compared to traditional methods using only two color channels and prior arts.

Bibliography

[1] U Rajendra Acharya, Hamido Fujita, Shu Lih Oh, Yuki Hagiwara, Jen Hong Tan, and Muhammad Adam. Application of Deep Convolutional Neural Network for Automated Detection of Myocardial Infarction Using ECG Signals. *Information Sciences*, 415:190–198, 2017.

[2] Michal Aharon, Michael Elad, Alfred Bruckstein, et al. K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation. *IEEE Transactions on Signal Processing*, 2006.

[3] Hanad Ahmed and Laurence Devoto. The potential of a digital twin in surgery. *Surgical Innovation*, 28(4):509–510, 2021.

[4] Nuzhat Ahmed and Yong Zhu. Early detection of atrial fibrillation based on ECG signals. *Bioengineering*, 7(1):16, 2020.

[5] Zia Uddin Ahmed, Mohammad Golam Mortuza, Mohammed Jashim Uddin, Md Humayun Kabir, Md Mahiuddin, and MD Jiabul Hoque. Internet of Things based patient health monitoring system using wearable biomedical device. In *2018 international conference on innovation in engineering and technology (ICIET)*, pages 1–5. IEEE, 2018.

[6] Mohammed Al-Disi, Hamza Djelouat, Christos Kotroni, Elena Politis, Abbes Amira, Faycal Bensaali, George Dimitrakopoulos, and Guillaume Alinier. ECG Signal Reconstruction on the IoT-gateway and Efficacy of Compressive Sensing Under Real-Time Constraints. *IEEE Access*, 2018.

[7] John Allen. Photoplethysmography and Its Application in Clinical Physiological Measurement. *Physiological Measurement*, 2007.

[8] Euan A Ashley and Josef Niebauer. *Cardiology explained*. Remedica, 2004.

[9] Md. Asif-Ur-Rahman, Fariha Afsana, Mufti Mahmud, M. Shamim Kaiser, Muhammad R. Ahmed, Omprakash Kaiwartya, and Anne James-Taylor. Toward a Heterogeneous Mist, Fog, and Cloud-Based Framework for the Internet of Healthcare Things. *IEEE Internet of Things J.*, 2019.

[10] Australian Radiation Protection and Nuclear Safety Agency. *Fitzpatrick skin phototype*.

[11] Teun Baar, Wiger van Houten, and Zeno Geradts. Camera identification by grouping images from database, based on shared noise patterns, 2012.

[12] Ufuk Bal. Non-contact Estimation of Heart Rate and Oxygen Saturation Using Ambient Light. *Biomed. Opt. Exp.*, Jan. 2015.

[13] Rohan Banerjee, Aniruddha Sinha, Anirban Dutta Choudhury, and Aishwarya Visvanathan. PhotoECG: Photoplethysmography to Estimate ECG Parameters. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014.

[14] Syed Khairul Bashar, Dong Han, Shirin Hajeb-Mohammadalipour, Eric Ding, Cody Whitcomb, David D McManus, and Ki H Chon. Atrial Fibrillation Detection from Wrist Photoplethysmography Signals Using Smartwatches. *Scientific reports*, 9(1):1–10, 2019.

[15] Dwaipayan Biswas, Luke Everson, Muqing Liu, Madhuri Panwar, Bram-Ernst Verhoef, Shrishail Patki, Chris H Kim, Amit Acharyya, Chris Van Hoof, Mario Konijnenburg, et al. Cornet: Deep learning framework for ppg-based heart rate estimation and biometric identification in ambulant environment. *IEEE transactions on biomedical circuits and systems*, 13(2):282–291, 2019.

[16] Koen Bruynseels, Filippo Santoni de Sio, and Jeroen van den Hoven. Digital Twins in Health Care: Ethical Implications of An Emerging Engineering Paradigm. *Frontiers in genetics*, 9:31, 2018.

[17] Nam Bui, Anh Nguyen, Phuc Nguyen, Hoang Truong, Ashwin Ashok, Thang Dinh, Robin Deterding, and Tam Vu. Smartphone-Based SpO2 Measurement by Exploiting Wavelengths Separation and Chromophore Compensation. *ACM Trans. Sens. Netw.*, Jan. 2020.

[18] Wilhelm Burger and Mark J. Burge. *Digital Image Processing - An Algorithmic Introduction using Java.* Springer, 2008.

[19] A John Camm. The role of continuous monitoring in atrial fibrillation management. *Arrhythmia & Electrophysiology Review*, 3(1):48, 2014.

[20] Cardiovascular diseases (CVDs). https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds). Accessed: 2022-06-09.

[21] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1721–1730, 2015.

[22] Gabriella Casalino, Giovanna Castellano, and Gianluca Zaza. A mHealth solution for contact-less self-monitoring of blood oxygen saturation. In *IEEE Symposium on Computers and Communications (ISCC)*, Jul. 2020.

[23] Douglas Chai and King N Ngan. Face segmentation using skin-color map in videophone applications. *IEEE Trans. Circuits and Systems for Video Technology*, 9(4):551–564, Jun. 1999.

[24] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A Library for Support Vector Machines. *ACM Trans. Intelligent Systems and Technology*, May 2011.

[25] Mingliang Chen, Qiang Zhu, Min Wu, and Quanzeng Wang. Modulation model of the photoplethysmography signal for vital sign extraction. *IEEE Journal of Biomedical and Health Informatics*, Aug. 2020.

[26] Weixuan Chen and Daniel McDuff. DeepPhys: Video-based physiological measurement using convolutional attention networks. In *The European Conference on Computer Vision (ECCV)*, pages 349–365, 2018.

[27] Yang Chen, Joo Heung Yoon, Michael R Pinsky, Ting Ma, and Gilles Clermont. Development of hemorrhage identification model using non-invasive vital signs. *Physiological measurement*, 41(5):055010, 2020.

[28] Hong-Yu Chiu, Hong-Han Shuai, and Paul C.-P. Chao. Reconstructing qrs complex from ppg by transformed attentional neural networks. *IEEE Sensors Journal*, 20(20):12374–12383, 2020.

[29] Youngjun Cho, Nadia Bianchi-Berthouze, and Simon J Julier. Deepbreath: Deep learning of breathing patterns for automatic stress recognition using low-cost thermal imaging in unconstrained settings. In *2017 seventh international conference on affective computing and intelligent interaction (acii)*, pages 456–463. IEEE, 2017.

[30] Eric Chern-Pin Chua, Stephen J Redmond, Gary McDarby, and Conor Heneghan. Towards Using Photo-plethysmogram Amplitude to Measure Blood Pressure During Sleep. *Annals of Biomedical Engineering*, 2010.

[31] Charles J Coté, E Andrew Goldstein, William H Fuchsman, and David C Hoaglin. The effect of nail polish on pulse oximetry. *Anesthesia and analgesia*, Jul. 1988.

[32] Jennifer Couzin-Frankel. The Mystery of The Pandemic's 'Happy Hypoxia'. *Science*, 2020.

[33] Darren Craven, Brian McGinley, Liam Kilmartin, Martin Glavin, and Edward Jones. Adaptive Dictionary Reconstruction for Compressed Sensing of ECG Signals. *IEEE Journal of Biomedical and Health Informatics*, 2016.

[34] Gerard De Haan and Vincent Jeanne. Robust pulse rate from chrominance-based rppg. *IEEE Transactions on Biomedical Engineering*, Jun. 2013.

[35] Anneke de Torbal, Eric Boersma, Jan A Kors, Gerard van Herpen, Jaap W Deckers, Deirdre AM van der Kuip, Bruno H Stricker, Albert Hofman, and Jacqueline CM Witteman. Incidence of recognized and unrecognized myocardial infarction in men and women aged 55 and older: the rotterdam study. *European heart journal*, Mar. 2006.

[36] Diagnose your irregular heart rhythm faster and more reliably with Zio. https://www.irhythmtech.com/patients/how-it-works. Accessed: 2022-06-17.

[37] Xinyi Ding, Damoun Nassehi, and Eric C Larson. Measuring Oxygen Saturation With Smartphone Cameras Using Convolutional Neural Networks. *IEEE Journal of Biomed. Health Informat.*, Dec. 2018.

[38] Carl Doersch. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*, 2016.

[39] ECG changes due to electrolyte imbalance (disorder). https://ecgwaves.com/topic/ecg-electrolyte-imbalance-electrolyte-disorder-calcium-potassium-magnesium/. Accessed: 2022-07-06.

[40] Empetica care: Unlock better health for thousands. https://www.empatica.com/care/. Accessed: 2022-07-14.

[41] Kjersti Engan, Sven Ole Aase, and J Hakon Husoy. Method of Optimal Directions for Frame Design. In *IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (ICASSP)*, 1999.

[42] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639):115–118, 2017.

[43] Martin Faulhaber, Hannes Gatterer, Thomas Haider, Tobias Linser, Nikolaus Netzer, and Martin Burtscher. Heart rate and blood pressure responses during hypoxic cycles of a 3-week intermittent hypoxia breathing program in patients at risk for or with mild copd. *International Journal of Chronic Obstructive Pulmonary Disease*, 2015.

[44] Riccardo Favilla, Veronica Chiara Zuccala, and Giuseppe Coppini. Heart rate and heart rate variability from single-channel video and ica integration of multiple signals. *IEEE journal of biomedical and health informatics*, Nov. 2018.

[45] Aidan Fuller, Zhong Fan, Charles Day, and Chris Barlow. Digital twin: Enabling technologies, challenges and open research. *IEEE access*, 8:108952–108971, 2020.

[46] W Bruce Fye. A history of the origin, evolution, and impact of electrocardiography. *The American journal of cardiology*, 73(13):937–949, 1994.

[47] Eduardo Gil, Michele Orini, Raquel Bailon, José María Vergara, Luca Mainardi, and Pablo Laguna. Photoplethysmography Pulse Rate Variability as A Surrogate Measurement of Heart Rate Variability During Non-stationary Conditions. *Physiological measurement*, 2010.

[48] Edward Glaessgen and David Stargel. The digital twin paradigm for future NASA and US Air Force vehicles. In *53rd AIAA/ASME/ASCE/AHS/ASC structures, structural dynamics and materials conference 20th AIAA/ASME/AHS adaptive structures conference 14th AIAA*, page 1818, 2012.

[49] Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of A New Research Resource for Complex Physiologic Signals. *Circulation*, 2000.

[50] B.W. Green, P. J.; Silverman. *Nonparametric Regression and Generalized Linear Models*. Chapman and Hall, 1990.

[51] Michael Grieves. Digital twin: manufacturing excellence through virtual factory replication. *White paper*, 1:1–7, 2014.

[52] Michael Grieves and John Vickers. *Digital twin: Mitigating unpredictable, undesirable emergent behavior in complex systems*, pages 85–113. Springer, 2017.

[53] Albinas Grunovas, Eugenijus Trinkunas, Alfonsas Buliuolis, Eurelija Venskaityte, and Jonas Poderys. Cardiovascular response to breath-holding explained by changes of the indices and their dynamic interactions. *Biological Systems: Open Access*, 2016.

[54] Alessandro R Guazzi, Mauricio Villarroel, Joao Jorge, Jonathan Daly, Matthew C Frise, Peter A Robbins, and Lionel Tarassenko. Non-contact Measurement of Oxygen Saturation with An RGB Camera. *Biomed. Opt. Express*, Sep. 2015.

[55] Hadi Habibzadeh, Karthik Dinesh, Omid Rajabi Shishvan, Andrew Boggio-Dandry, Gaurav Sharma, and Tolga Soyata. A Survey of Healthcare Internet of Things (HIoT): A Clinical Perspective. *IEEE Internet of Things J.*, 2020.

[56] Adi Hajj-Ahmad, Ravi Garg, and Min Wu. Instantaneous frequency estimation and localization for enf signals. In *Proc. 4th Annu. Summit and Conf. (APSIPA)*. IEEE, Dec. 2012.

[57] John Hampton and Joanna Hampton. *The ECG Made Easy E-Book*. Elsevier Health Sciences, 2019.

[58] Awni Y Hannun, Pranav Rajpurkar, Masoumeh Haghpanahi, Geoffrey H Tison, Codie Bourn, Mintu P Turakhia, and Andrew Y Ng. Cardiologist-level Arrhythmia Detection and Classification in Ambulatory Electrocardiograms Using a Deep Neural Network. *Nature medicine*, 25(1):65–69, 2019.

[59] Awni Y Hannun, Pranav Rajpurkar, Masoumeh Haghpanahi, Geoffrey H Tison, Codie Bourn, Mintu P Turakhia, and Andrew Y Ng. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature medicine*, 25(1):65–69, 2019.

[60] Simon S Haykin. *Adaptive filter theory*. Pearson Education India, 2008.

[61] Lara J. Herbert and Iain H. Wilson. Pulse oximetry in low-resource settings. *Breathe*, 9(2):90–98, 2012.

[62] Jason S Hoffman, Varun Viswanath, Xinyi Ding, Matthew J Thompson, Eric C Larson, Shwetak N Patel, and Edward Wang. Smartphone camera oximetry in an induced hypoxemia study. *arXiv preprint arXiv:2104.00038*, 2021.

[63] Holter monitor. https://www.hopkinsmedicine.org/health/treatment-tests-and-therapies/holter-monitor. Accessed: 2022-07-14.

[64] How to use the Blood Oxygen app on Apple Watch Series 6. https://support.apple.com/en-us/HT211027. Accessed: 2021-05-17.

[65] International Organization for Standardization. *Particular requirements for basic safety and essential performance of pulse oximeter equipment* , 2011.

[66] Luca Iozzia, Luca Cerina, and Luca Mainardi. Relationships between heart-rate variability and pulse-rate variability obtained from video-ppg signal using zca. *Physiological measurement*, Sep. 2016.

[67] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning*. Springer, 2013.

[68] In Cheol Jeong and Joseph Finkelstein. Introducing contactless blood pressure assessment using a high speed video camera. *Journal of medical systems*, Apr. 2016.

[69] Zhuolin Jiang, Zhe Lin, and Larry S Davis. Label Consistent K-SVD: Learning a Discriminative Dictionary for Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013.

[70] Anders Johansson. Neural Network for Photoplethysmographic Respiratory Rate Monitoring. *Medical and Biological Engineering and Computing*, 2003.

[71] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. MIMIC-III, A Freely Accessible Critical Care Database. *Scientific Data*, 2016.

[72] Anand Kumar Joshi, Arun Tomar, and Mangesh Tomar. A Review Paper on Analysis of Electrocardiograph (ECG) Signal for the Detection of Arrhythmia Abnormalities. *International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering*, 2014.

[73] KardiaMobile: Check in on your heart from home. https://store.kardia.com/products/kardiamobile. Accessed: 2022-1-25.

[74] Walter Karlen, Srinivas Raman, J Mark Ansermino, and Guy A Dumont. Multiparameter Respiratory Rate Estimation from the Photoplethysmogram. *IEEE Trans. Biomed. Eng.*, 60(7):1946–1953, 2013.

[75] Emroz Khan, Forsad Al Hossain, Shiekh Zia Uddin, S Kaisar Alam, and Md Kamrul Hasan. A Robust Heart Rate Monitoring Scheme Using Photoplethysmographic Signals Corrupted by Intense Motion Artifacts. *IEEE Transactions on Biomedical Engineering*, 63(3):550–562, 2016.

[76] Paul Kligfield, Leonard S Gettes, James J Bailey, Rory Childers, Barbara J Deal, E William Hancock, Gerard Van Herpen, Jan A Kors, Peter Macfarlane, David M Mirvis, et al. Recommendations for the standardization and interpretation of the electrocardiogram: part i: the electrocardiogram and its technology a scientific statement from the american heart association electrocardiography and arrhythmias committee, council on clinical cardiology; the american college of cardiology foundation; and the heart rhythm society endorsed by the international society for computerized electrocardiology. *Journal of the American College of Cardiology*, 49(10):1109–1127, 2007.

[77] Lingqin Kong, Yuejin Zhao, Liquan Dong, Yiyun Jian, Xiaoli Jin, Bing Li, Yun Feng, Ming Liu, Xiaohua Liu, and Hong Wu. Non-contact detection of oxygen saturation based on visible light imaging device using ambient light. *Opt. Exp.*, Jul. 2013.

[78] John K Kruschke. Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, 142(2):573, 2013.

[79] John K Kruschke. Rejecting or accepting parameter values in bayesian estimation. *Advances in Methods and Practices in Psychological Science*, 2018.

[80] John K Kruschke. Bayesian analysis reporting guidelines. *Nature Human Behaviour*, pages 1–10, 2021.

[81] John K Kruschke and Torrin M Liddell. Bayesian data analysis for newcomers. *Psychonomic bulletin & review*, 25(1):155–177, 2018.

[82] John K Kruschke and Torrin M Liddell. The bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a bayesian perspective. *Psychonomic Bulletin & Review*, 25(1):178–206, 2018.

[83] Aparna Kumari, Sudeep Tanwar, Sudhanshu Tyagi, and Neeraj Kumar. Fog computing for Healthcare 4.0 environment: Opportunities and challenges. *Computers and Electrical Engineering*, 72:1–13, 2018.

[84] Jingshan Li and Pascale Carayon. Health Care 4.0: A vision for smart and connected health care. *IISE transactions on healthcare systems engineering*, 11(3):171–180, 2021.

[85] Kai Li, Zhengming Ding, Sheng Li, and Yun Fu. Discriminative Semi-coupled Projective Dictionary Learning for Low-resolution Person Re-identification. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[86] Lisha Li, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. Hyperband: A novel bandit-based approach to hyperparameter optimization. *The Journal of Machine Learning Research*, Apr. 2018.

[87] Xiaobai Li, Jie Chen, Guoying Zhao, and Matti Pietikainen. Remote heart rate measurement from face videos under realistic situations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4264–4271, 2014.

[88] Yuenan Li, Xin Tian, Qiang Zhu, and Min Wu. A lightweight neural network for inferring ecg and diagnosing cardiovascular diseases from ppg. *arXiv preprint arXiv:2012.04949*, 2020.

[89] Zhicheng Li, Hong Huang, and Satyajayant Misra. Compressed Sensing via Dictionary Learning and Approximate Message Passing for Multimedia Internet of Things. *IEEE Internet of Things J.*, 2017.

[90] Tong Liu, Yujuan Si, Dunwei Wen, Mujun Zang, and Liuqi Lang. Dictionary Learning for VQ Feature Extraction in ECG Beats Classification. *Expert Systems with Applications*, 2016.

[91] Ying Liu, Lin Zhang, Yuan Yang, Longfei Zhou, Lei Ren, Fei Wang, Rong Liu, Zhibo Pang, and M Jamal Deen. A novel cloud-based framework for the elderly healthcare services using digital twin. *IEEE Access*, 7:49088–49101, 2019.

[92] Zhiyuan Lu, Xiang Chen, Zhongfei Dong, Zhangyan Zhao, and Xu Zhang. A Prototype of Reflection Pulse Oximeter Designed for Mobile Healthcare. *IEEE Journal of Biomed. Health Informat.*, Aug. 2015.

[93] Julien Mairal, Jean Ponce, Guillermo Sapiro, Andrew Zisserman, and Francis R Bach. Supervised Dictionary Learning. In *Advances in Neural Information Processing Systems*, 2009.

[94] Angshul Majumdar and Rabab Ward. Robust Greedy Deep Dictionary Learning for ECG Arrhythmia Classification. In *IEEE International Joint Conference on Neural Networks (IJCNN)*, 2017.

[95] Daniel McDuff, Sarah Gontarek, and Rosalind Picard. Remote measurement of cognitive stress via heart rate variability. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 2957–2960. IEEE, 2014.

[96] Yunyoung Nam, Bersain A Reyes, and Ki H Chon. Estimation of respiratory rates using the built-in microphone of a smartphone or headset. *IEEE Journal of Biomedical and Health Informatics*, Sep. 2015.

[97] Angela Navarrete-Opazo and Gordon S Mitchell. Therapeutic potential of intermittent hypoxia: a matter of dose. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, 307(10):R1181–R1197, 2014.

[98] Andrea Nemcova, Ivana Jordanova, Martin Varecka, Radovan Smiseka, Lucie Marsanova, Lukas Smital, and Martin Vitek. Monitoring of heart rate, blood oxygen saturation, and blood pressureusing a smartphone. *Biomedical Signal Processing and Control*, May 2020.

[99] Masataka Nishiga, Dao Wen Wang, Yaling Han, David B Lewis, and Joseph C Wu. COVID-19 and cardiovascular disease: from basic mechanisms to clinical perspectives. *Nature Reviews Cardiology*, 17(9):543–558, 2020.

[100] Meir Nitzan, Ayal Romem, and Robert Koppel. Pulse oximetry: Fundamentals and technology update. *Medical Devices (Auckland, NZ)*, 7:231, 2014.

[101] Xuesong Niu, Shiguang Shan, Hu Han, and Xilin Chen. RhythmNet: End-to-end heart rate estimation from face via spatial-temporal representation. *IEEE Trans. on Image Processing*, Oct. 2019.

[102] Optical Absorption of Hemoglobin. https://omlc.org/spectra/hemoglobin/. Accessed: 2021-03-09.

[103] Nobuyuki Otsu. A threshold Selection Method from Gray-level Histograms. *IEEE Trans. Syst., Man, and Cybernet.*, Jan. 1979.

[104] Jiapu Pan and Willis J. Tompkins. A Real-time QRS Detection Algorithm. *IEEE Transactions on Biomedical Engineering*, 1985.

[105] Neeraj Paradkar and Shubhajit Roy Chowdhury. Cardiac Arrhythmia Detection Using Photoplethysmography. In *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 113–116. IEEE, 2017.

[106] Judea Pearl. *Causality*. Cambridge university press, 2009.

[107] Marco A. F. Pimentel, Alistair E. W. Johnson, Peter H. Charlton, Drew Birrenkott, Peter J. Watkinson, Lionel Tarassenko, and David A. Clifton. Toward a robust estimation of respiratory rate from pulse oximeters. *IEEE Transactions on Biomedical Engineering*, 64(8):1914–1923, 2017.

[108] Annette Plüddemann, Matthew Thompson, Carl Heneghan, and Christopher Price. Pulse Oximetry in Primary Care: Primary Care Diagnostic Technology Update. *British Journal of General Practice*, May 2011.

[109] Ming-Zher Poh, Daniel J McDuff, and Rosalind W Picard. Advancements in noncontact, multiparameter physiological measurements using a webcam. *IEEE transactions on biomedical engineering*, Oct. 2010.

[110] R Package for BEST: Bayesian Estimation Supersedes the t-Test. https://CRAN.R-project.org/package=BEST. Accessed: 2021-09-30.

[111] Natasa Reljin, Gary Zimmer, Yelena Malyuta, Yitzhak Mendelson, Chad E Darling, and Ki H Chon. Detection of blood loss in trauma patients using time-frequency analysis of photoplethysmographic signal. In *2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, pages 118–121. IEEE, 2016.

[112] Natasa Reljin, Gary Zimmer, Yelena Malyuta, Kirk Shelley, Yitzhak Mendelson, David J Blehar, Chad E Darling, and Ki H Chon. Using support vector machines on photoplethysmographic signals to discriminate between hypovolemia and euvolemia. *PLoS One*, 13(3):e0195087, 2018.

[113] Alessandra Rosa and Roberto Cesar Betini. Noncontact SpO2 Measurement Using Eulerian Video Magnification. *IEEE Trans. Instrum. Meas.*, May 2019.

[114] Anna Rosiek and Krzysztof Leksowski. The Risk Factors and Prevention of Cardiovascular Disease: The Importance of Electrocardiogram in the Diagnosis and Treatment of Acute Coronary Syndrome. *Therapeutics and Clinical Risk Management*, 2016.

[115] Gregory A Roth, George A Mensah, Catherine O Johnson, Giovanni Addolorato, Enrico Ammirati, Larry M Baddour, Noël C Barengo, Andrea Z Beaton, Emelia J Benjamin, and Catherine P Benziger. Global burden of cardiovascular diseases and risk factors, 1990–2019: update from the GBD 2019 study. *Journal of the American College of Cardiology*, 76(25):2982–3021, 2020.

[116] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

[117] Christopher G Scully, Jinseok Lee, Joseph Meyer, Alexander M Gorbach, Domhnull Granquist-Fraser, Yitzhak Mendelson, and Ki H Chon. Physiological Parameter Monitoring from Optical Recordings with A Mobile Phone. *IEEE Trans. Biomed. Eng.*, Jul. 2011.

[118] Hooman Sedghamiz. BioSigKit: A Matlab Toolbox and Interface for Analysis of BioSignals. *Journal of Open Source Software*, 2018.

[119] Hooman Sedghamiz and Daniele Santonocito. Unsupervised Detection and Classification of Motor Unit Action Potentials in Intramuscular Electromyography Signals. In *IEEE E-health and Bioengineering Conference (EHB)*, 2015.

[120] Servier Medical Art. https://smart.servier.com/?s=heart. Accessed: 2022-07-14.

[121] John W Severinghaus. Takuo Aoyagi: Discovery of pulse oximetry. *Anesthesia & Analgesia*, Dec. 2007.

[122] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

[123] Dangdang Shao, Chenbin Liu, Francis Tsow, Yuting Yang, Zijian Du, Rafael Iriya, Hui Yu, and Nongjian Tao. Noncontact Monitoring of Blood Oxygen Saturation Using Camera and Dual-wavelength Imaging System. *IEEE Trans. Biomed. Eng*, Sep. 2015.

[124] Niraj Shenoy, Rebecca Luchtel, and Perminder Gulani. Considerations for Target Oxygen Saturation in COVID-19 Patients: Are We Under-shooting? *BMC Medicine*, Dec. 2020.

[125] M. Celeste Simon and Brian Keith. The Role of Oxygen Availability in Embryonic Development and Stem Cell Function. *Nature Reviews Molecular Cell Biology*, Apr. 2008.

[126] Kwanghyun Sohn, Faisal M Merchant, Omid Sayadi, Dheeraj Puppala, Rajiv Doddamani, Ashish Sahani, Jagmeet P Singh, E Kevin Heist, Eric M Isselbacher, and Antonis A Armoundas. A novel point-of-care smartphone based system for monitoring the cardiac and respiratory systems. *Scientific Reports*, Mar. 2017.

[127] Radim Špetlík, Vojtech Franc, and Jirí Matas. Visual heart rate estimation with convolutional neural network. In *British Machine Vision Conf., Newcastle, UK*, Sep. 2018.

[128] Steven R Steinhubl, Jill Waalen, Alison M Edwards, Lauren M Ariniello, Rajesh R Mehta, Gail S Ebner, Chureen Carter, Katie Baca-Motes, Elise Felicione, Troy Sarich, et al. Effect of a home-based wearable continuous ecg monitoring patch on detection of undiagnosed atrial fibrillation: the mstops randomized clinical trial. *Jama*, 320(2):146–155, 2018.

[129] Yu Sun and Nitish Thakor. Photoplethysmography revisited: from contact to noncontact, from point to imaging. *IEEE transactions on biomedical engineering*, Sep. 2015.

[130] Zhiyuan Sun, Qinghua He, Yuandong Li, Wendy Wang, and Ruikang K Wang. Robust non-contact peripheral oxygenation saturation measurement using smartphone-enabled imaging photoplethysmography. *Biomed. Opt. Exp.*, 12(3):1746–1760, Mar. 2021.

[131] M Suresh and Urmila Natarajan. Healthcare 4.0: Recent advances and futuristic research avenues. *Materials Today: Proceedings*, 2021.

[132] Take an ECG with the ECG app on Apple Watch. https://support.apple.com/en-us/HT208955. Accessed: 2022-1-25.

[133] Lionel Tarassenko, Mauricio Villarroel, Alessandro Guazzi, João Jorge, DA Clifton, and Chris Pugh. Non-contact Video-based Vital Sign Monitoring Using Ambient Light and Auto-regressive Models. *Physiol. Meas*, Mar. 2014.

[134] İsmail Tayfur and Mustafa Ahmet Afacan. Reliability of smartphone measurements of vital parameters: A prospective study using a reference method. *The American J. Emergency Medicine*, 37(8):1527–1530, Aug. 2019.

[135] Jason Teo. Early Detection of Silent Hypoxia in COVID-19 Pneumonia Using Smartphone Pulse Oximetry. *Journal of Medical Systems*, Aug. 2020.

[136] Xin Tian, Qiang Zhu, Yuenan Li, and Min Wu. Cross-Domain Joint Dictionary Learning for ECG Reconstruction from PPG. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 936–940, 2020.

[137] Xin Tian, Qiang Zhu, Yuenan Li, and Min Wu. Cross-domain Joint Dictionary Learning for ECG Inference from PPG. *arXiv preprint arXiv:2101.02362*, 2021. Under review.

[138] Martin J Tobin, Franco Laghi, and Amal Jubran. Why COVID-19 Silent Hypoxemia is Baffling to Physicians. *American Journal of Respiratory and Critical Care Medicine*, Aug. 2020.

[139] Joel A Tropp and Anna C Gilbert. Signal Recovery from Random Measurements via Orthogonal Matching Pursuit. *IEEE Transactions on Information Theory*, 2007.

[140] Hsin-Yi Tsai, Kuo-Cheng Huang, and J Andrew Yeh. No-contact oxygen saturation measuring technology for skin tissue and its application. *IEEE Instrum. Meas. Magazine*, Sep. 2016.

[141] Sergey Tulyakov, Xavier Alameda-Pineda, Elisa Ricci, Lijun Yin, Jeffrey F Cohn, and Nicu Sebe. Self-adaptive matrix completion for heart rate estimation from face videos under realistic conditions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2396–2404, 2016.

[142] Mark Van Gastel, Sander Stuijk, and Gerard De Haan. New Principle for Measuring Arterial Blood Oxygenation, Enabling Motion-Robust Remote Monitoring. *Scientific Reports*, Dec. 2016.

[143] Mark van Gastel, Wim Verkruysse, and Gerard de Haan. Data-driven Calibration Estimation for Robust Remote Pulse-oximetry. *Applied Sciences*, Jan. 2019.

[144] JP Varshney. *Electrocardiography in Veterinary Medicine*. Springer, 2020.

[145] Wim Verkruysse, Lars O Svaasand, and J Stuart Nelson. Remote plethysmographic imaging using ambient light. *Opt. Exp.*, Dec. 2008.

[146] Adriana N Vest, Giulia Da Poian, Qiao Li, Chengyu Liu, Shamim Nemati, Amit J Shah, and Gari D Clifford. An Open Source Benchmarked Toolbox for Cardiovascular Waveform and Interval Analysis. *Physiological Measurement*, 2018.

[147] Khuong Vo, Emad Kasaeyan Naeini, Amir Naderi, Daniel Jilani, Amir M Rahmani, Nikil Dutt, and Hung Cao. P2E-WGAN: ECG waveform synthesis from PPG with conditional wasserstein generative adversarial networks. In *Proceedings of the 36th Annual ACM Symposium on Applied Computing*, pages 1030–1036, 2021.

[148] Shenlong Wang, Lei Zhang, Yan Liang, and Quan Pan. Semi-coupled Dictionary Learning with Applications to Image Super-resolution and Photo-sketch Synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[149] Wenjin Wang, Albertus C den Brinker, Sander Stuijk, and Gerard De Haan. Algorithmic principles of remote PPG. *IEEE Trans. on Biomedical Eng.*, Sep. 2016.

[150] Wenjin Wang, Sander Stuijk, and Gerard De Haan. A novel algorithm for remote photoplethysmography: Spatial subspace rotation. *IEEE transactions on biomedical engineering*, Dec. 2015.

[151] Larry Wasserman. *All of statistics: a concise course in statistical inference*, volume 26. Springer, 2004.

[152] John G Webster. *Design of Pulse Oximeters*. CRC Press, Oct. 1997.

[153] Taiyang Wu, Fan Wu, Chunkai Qiu, Jean-Michel Redouté, and Mehmet Rasit Yuce. A Rigid-Flex Wearable Health Monitoring Sensor Patch for IoT-Connected Healthcare Applications. *IEEE Internet of Things Journal*, 2020.

[154] Jian Xu, Chun Qi, and Zhiguo Chang. Coupled K-SVD Dictionary Training for Super-resolution. In *IEEE International Conference on Image Processing (ICIP)*, 2014.

[155] Jianchao Yang, Zhaowen Wang, Zhe Lin, Scott Cohen, and Thomas Huang. Coupled Dictionary Training for Image Super-resolution. *IEEE Transactions on Image Processing*, 2012.

[156] Jianchao Yang, John Wright, Thomas S Huang, and Yi Ma. Image Super-resolution via Sparse Representation. *IEEE Transactions on Image Processing*, 2010.

[157] Gülendam Hakverdioğlu Yönt, Esra Akin Korhan, and Berna Dizer. The effect of nail polish on pulse oximetry readings. *Intensive and Critical Care Nursing*, Apr. 2014.

[158] Yue Yu, Jie Chen, Tian Gao, and Mo Yu. DAG-GNN: DAG structure learning with graph neural networks. In *International Conference on Machine Learning*, pages 7154–7163. PMLR, 2019.

[159] Gaobo Zhang, Zhen Mei, Yuan Zhang, Xuesheng Ma, Benny Lo, Dongyi Chen, and Yuanting Zhang. A Noninvasive Blood Glucose Monitoring System Based on Smartphone PPG Signal Processing and Machine Learning. *IEEE Transactions on Industrial Informatics*, 16(11):7209–7218, 2020.

[160] Zheng Zhang, Yong Xu, Jian Yang, Xuelong Li, and David Zhang. A Survey of Sparse Representation: Algorithms and Applications. *IEEE Access*, 2015.

[161] Zhilin Zhang, Zhouyue Pi, and Benyuan Liu. TROIKA: A General Framework for Heart Rate Monitoring Using Wrist-type Photoplethysmographic Signals During Intensive Physical Exercise. *IEEE Trans. Biomed. Eng.*, 2014.

[162] Yufeng Zheng, Hongyu Wang, and Yingguang Hao. Mobile application for monitoring body temperature from facial images using convolutional neural network and support vector machine. *Mobile Multimedia/Image Processing, Security, and Applications*, April 2020.

[163] Qiang Zhu, Mingliang Chen, Chau-Wai Wong, and Min Wu. Adaptive Multi-Trace Carving for Robust Frequency Tracking in Forensic Applications. *IEEE Trans. Inf. Forensics Security*, May 2020.

[164] Qiang Zhu, Xin Tian, Chau-Wai Wong, and Min Wu. ECG Reconstruction via PPG: A Pilot Study. In *IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, Chicago, IL, May 2019.

[165] Qiang Zhu, Xin Tian, Chau-Wai Wong, and Min Wu. Learning Your Heart Actions From Pulse: ECG Waveform Reconstruction From PPG. *IEEE Internet of Things Journal*, 8(23):16734–16748, 2021.

[166] Qiang Zhu, Chau-Wai Wong, Chang-Hong Fu, and Min Wu. Fitness heart rate measurement using face videos. In *IEEE International Conference on Image Processing (ICIP)*, Sep. 2017.