UNIVERSITY OF CALIFORNIA SAN DIEGO

Machine Learning Techniques for Personalized Health Monitoring and Interventions using Wearable Device Data

A Dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy

in

Electrical Engineering (Intelligent Systems, Robotics, and Control)

by

Jared Leitner

Committee in charge:

Professor Sujit Dey, Chair
Professor Ramesh Rao, Co-Chair
Professor Peter Gerstoft
Professor Loki Natarajan
Professor Edward Wang

2024

The Dissertation of Jared Leitner is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2024

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I extend my sincere appreciation to my PhD advisor, Prof. Sujit Dey, for his guidance and support throughout my doctoral studies. I am also grateful to my dissertation committee members, Prof. Ramesh Rao, Prof. Peter Gerstoft, Prof. Loki Natarajan, and Prof. Edward Wang, for their valuable insights and constructive feedback.

I would also like to thank the collaborators who worked closely with me throughout my research: Dr. Po-Han Chiang, Dr. Brian Khan, and Dr. Parag Agnihotri. Their collaboration and expertise have greatly enhanced the quality of my work and made my research experience rewarding.

Lastly, I express my deepest gratitude to my parents, sister, and girlfriend, who have been a constant source of support. Their encouragement, patience, and love have been invaluable throughout this journey.

Chapter 1, in part, is from the material as it appears in the IEEE Journal of Biomedical and Health Informatics, 2022, Leitner, Jared; Chiang, Po-Han; Dey, Sujit. The dissertation author was the primary investigator and author of this paper.

Chapter 2, in part, is from the material as it appears in the IEEE Journal of Biomedical and Health Informatics, 2023, Leitner, Jared; Alexander, Behnke; Chiang, Po-Han; Ritter, Michele; Millen, Marlene; Dey, Sujit. The dissertation author was the primary investigator and author of this paper.

Chapter 3, in part, is from the material as it appears in the Journal of Medical Internet Research, Cardio, 2024, Leitner, Jared; Chiang, Po-Han; Agnihotri, Parag; Dey, Sujit. The dissertation author was the primary investigator and author of this paper.

VITA

2018 Bachelor of Science in Electrical Engineering, University of California Los Angeles

2020 Master of Science in Electrical Engineering, University of California San Diego

2024 Doctor of Philosophy in Electrical Engineering, University of California San Diego

PUBLICATIONS

J. Leitner, P. -H. Chiang and S. Dey, "Personalized Blood Pressure Estimation using Photoplethysmography and Wavelet Decomposition," 2019 IEEE International Conference on E-health Networking, Application & Services (HealthCom), Bogota, Colombia, 2019, pp. 1-6, doi: 10.1109/HealthCom46333.2019.9009587.

J. Leitner, P. -H. Chiang and S. Dey, "Personalized Blood Pressure Estimation Using Photoplethysmography: A Transfer Learning Approach," in IEEE Journal of Biomedical and Health Informatics, vol. 26, no. 1, pp. 218-228, Jan. 2022, doi: 10.1109/JBHI.2021.3085526.

J. Leitner, P. -H. Chiang, B. Khan and S. Dey, "An mHealth Lifestyle Intervention Service for Improving Blood Pressure using Machine Learning and IoMTs," 2022 IEEE International Conference on Digital Health (ICDH), Barcelona, Spain, 2022, pp. 142-150, doi: 10.1109/ICDH55609.2022.00030.

J. Leitner, A. Behnke, P. -H. Chiang, M. Ritter, M. Millen and S. Dey, "Classification of Patient Recovery From COVID-19 Symptoms Using Consumer Wearables and Machine Learning," in IEEE Journal of Biomedical and Health Informatics, vol. 27, no. 3, pp. 1271-1282, March 2023, doi: 10.1109/JBHI.2023.3239366.

J. Leitner, P. -H. Chiang, P. Agnihotri and S. Dey, “The Effect of an AI-Based, Autonomous, Digital Health Intervention Using Precise Lifestyle Guidance on Blood Pressure in Adults With Hypertension: Single-Arm Nonrandomized Trial,” in JMIR Cardio, May 2024. doi:10.2196/51916.

S. Xie, J. Leitner and S. Dey, "Personalized Impact of Lifestyle on Type 1 Diabetes Patients: A Comprehensive Regression Analysis," 2024 IEEE International Conference on Healthcare Informatics, Orlando, Florida, USA, 2024.

ABSTRACT OF THE DISSERTATION

Machine Learning Techniques for Personalized Health Monitoring
and Interventions using Wearable Device Data

by

Jared Leitner

Doctor of Philosophy in Electrical Engineering (Intelligent Systems, Robotics, and Control)

University of California San Diego, 2024

Professor Sujit Dey, Chair
Professor Ramesh Rao, Co-Chair

To provide more optimal care at scale, health systems are changing the way in which healthcare is delivered. At the center of this changing landscape is a shift towards remote, continuous, and automated delivery of healthcare. This shift can lead to significant improvement in and scalability of at-home patient care for chronic diseases like hypertension and viral illnesses like COVID-19, while at the same time enabling significant savings in human and equipment resources. Wearable devices are an enabling technology making this shift in healthcare delivery

possible due to the substantial amount of lifestyle and vitals data they can remotely collect. There is great opportunity for machine learning (ML) to assist in the remote and personalized delivery of care due to the large amount of data that is collected.

In this dissertation, we present three applications of ML to enable personalized, remote health monitoring and care delivery. Chapter 1 presents a personalized deep learning approach to estimate blood pressure (BP) using the photoplethysmogram signal. Our approach enables continuous, noninvasive BP monitoring as compared to traditional methods which are either intermittent or invasive. To address the problem of limited personal data for individuals, we propose a transfer learning technique that achieves a mean absolute error of 3.52 and 2.20 mmHg for systolic and diastolic BP estimation, respectively. Chapter 2 describes a ML-based remote monitoring method to estimate patient recovery from COVID-19 symptoms using automatically collected wearable device data, instead of relying on manually collected symptom data. Our method achieves an F1-score of 0.88 when applying our Random Forest-based model personalization technique using weighted bootstrap aggregation. Chapter 3 presents the results of a single-arm nonrandomized trial which assessed the effectiveness of a fully digital, autonomous, and ML–based lifestyle coaching program on achieving BP control among adults with hypertension. 141 participants were monitored over 24 weeks and achieved an average systolic and diastolic BP decrease of 8.1 mmHg and 5.1 mmHg, respectively. Our research demonstrates the successful application of ML across various healthcare contexts. By harnessing wearable device data, we can facilitate more personalized and effective monitoring and interventions.

INTRODUCTION

Healthcare systems are transitioning towards more remote, continuous, and automated healthcare delivery methods to provide more optimal care on a broader scale. This shift holds the promise of substantial improvements and scalability in at-home patient care, including chronic conditions like hypertension and infectious diseases such as COVID-19. Simultaneously, it offers the potential for significant savings in both human resources and equipment expenses. Wearable devices are a key technology enabling this shift in healthcare delivery, due to their ability to remotely gather a wealth of lifestyle and vital sign data. This influx of data presents a unique opportunity for machine learning (ML) to play a pivotal role in enabling remote and personalized care delivery. By harnessing the vast quantities of collected data, ML algorithms can enhance diagnostic accuracy, predict disease progression, and tailor interventions to individual patient contexts. In this dissertation, we present three applications of machine learning to enable enhanced personalized health monitoring and care delivery.

Chapter 1 presents a personalized deep learning approach to estimate blood pressure (BP) using the photoplethysmogram (PPG) signal. Our approach enables continuous, noninvasive BP monitoring as compared to traditional methods which are either intermittent or invasive. We propose a hybrid neural network architecture consisting of convolutional, recurrent, and fully connected layers that operates directly on the raw PPG time series and provides BP estimation every 5 seconds. To address the problem of limited personal PPG and BP data for individuals, we propose a transfer learning technique that personalizes specific layers of a network pre-trained with abundant data from other patients. We use the MIMIC III database which contains PPG and continuous BP data measured invasively via an arterial catheter to develop and analyze our approach. Our transfer learning technique, namely BP-CRNN-Transfer, achieves a mean absolute

error (MAE) of 3.52 and 2.20 mmHg for SBP and DBP estimation, respectively, outperforming existing methods. Our approach satisfies both the BHS and AAMI blood pressure measurement standards for SBP and DBP. Moreover, our results demonstrate that as little as 50 data samples per person are required to train accurate personalized models. We carry out Bland-Altman and correlation analysis to compare our method to the invasive arterial catheter, which is the gold-standard BP measurement method.

Chapter 2 discusses the limitations of current remote monitoring practices for COVID-19 patients, predominantly reliant on manual symptom reporting and patient compliance. Our contribution lies in proposing a machine learning (ML)-based remote monitoring approach that uses automatically gathered data from wearable devices. Our proposed method estimates patient recovery from COVID-19 symptoms, mitigating the dependence on manual symptom collection. We deploy our remote monitoring system, namely eCOVID, in two COVID-19 telemedicine clinics. Our system utilizes a Garmin wearable and symptom tracker mobile app for data collection. The data consists of vitals, lifestyle, and symptom information which is fused into an online report for clinicians to review. Symptom data collected via our mobile app is used to label the recovery status of each patient daily. We propose a ML-based binary patient recovery classifier which uses wearable data to estimate whether a patient has recovered from COVID-19 symptoms. We evaluate our method using leave-one-subject-out (LOSO) cross-validation, and find that Random Forest (RF) is the top performing model. Our method achieves an F1-score of 0.88 when applying our RF-based model personalization technique using weighted bootstrap aggregation. Our results demonstrate that ML-assisted remote COVID-19 monitoring using automatically collected wearable data can supplement or be used in place of manual daily symptom tracking which relies on patient compliance.

Chapter 3 presents the results of a single-arm nonrandomized trial which assessed the effectiveness of a fully digital, autonomous, and ML–based lifestyle coaching program on achieving BP control among adults with hypertension. Home BP monitoring with lifestyle coaching is effective in managing hypertension and reducing cardiovascular risk. However, traditional manual lifestyle coaching models significantly limit availability due to high operating costs and personnel requirements. Furthermore, the lack of patient lifestyle monitoring and clinician time constraints can prevent personalized coaching on lifestyle modifications. To address these challenges, we propose a ML-driven, autonomous, precise lifestyle coaching program for patients with hypertension. Participants who enrolled in the trial received a BP monitor and wearable activity tracker. Data were collected from these devices and a questionnaire mobile app and were used to train personalized ML models that enabled precision lifestyle coaching delivered to participants via SMS text messaging and a mobile app. The primary outcomes included (1) the changes in systolic and diastolic BP from baseline to 12 and 24 weeks and (2) the percentage change of participants in the controlled, stage 1, and stage 2 hypertension categories from baseline to 12 and 24 weeks. Secondary outcomes included (1) the participant engagement rate as measured by data collection consistency and (2) the number of manual clinician outreaches. In total, 141 participants were monitored over 24 weeks. At 12 weeks, systolic and diastolic BP decreased by 5.6 mm Hg (P<.001; 95% CI −7.1 to −4.2) and 3.8 mm Hg (P<.001; 95% CI −4.7 to −2.8), respectively. Particularly, for participants starting with stage 2 hypertension, systolic and diastolic BP decreased by 9.6 mm Hg (P<.001; 95% CI −12.2 to −6.9) and 5.7 mm Hg (P<.001; 95% CI −7.6 to −3.9), respectively. At 24 weeks, systolic and diastolic BP decreased by 8.1 mm Hg (P<.001; 95% CI −10.1 to −6.1) and 5.1 mm Hg (P<.001; 95% CI −6.2 to −3.9), respectively. For participants starting with stage 2 hypertension, systolic and diastolic BP decreased by 14.2 mm Hg

(*P*<.001; 95% CI −17.7 to −10.7) and 8.1 mm Hg (*P*<.001; 95% CI −10.4 to −5.7), respectively, at 24 weeks. The percentage of participants with controlled BP increased by 17.2% (22/128; *P*<.001) and 26.5% (27/102; *P*<.001) from baseline to 12 and 24 weeks, respectively. The percentage of participants with stage 2 hypertension decreased by 25% (32/128; *P*<.001) and 26.5% (27/102; *P*<.001) from baseline to 12 and 24 weeks, respectively. The average weekly participant engagement rate was 92% (SD 3.9%), and only 5.9% (6/102) of the participants required manual outreach over 24 weeks. The study demonstrates the potential of fully digital, autonomous, and ML–based lifestyle coaching to achieve meaningful BP improvements and high engagement for patients with hypertension, while substantially reducing clinician workloads.

Chapter 1 Personalized Blood Pressure Estimation Using Photoplethysmography: A Transfer Learning Approach

1.1 Introduction

Blood pressure (BP) is the most important indicator of cardiovascular health. High blood pressure, or hypertension, affects 30% of American adults and contributes to over 410,000 deaths per year [1,2]. This condition has been called "the silent killer," as typically no symptoms are recognized before significant damage has already been done to the heart and arteries [3]. BP is defined as the pressure exerted on the arteries as blood is pumped throughout the body and is measured in millimeters of mercury (mmHg). Systolic (SBP) and diastolic blood pressure (DBP) are the primary metrics used to measure BP, which are defined as the maximum and minimum blood pressure, respectively, during a pulse.

For accurate diagnosis and treatment of hypertension, regular BP measurement is necessary. According to the American College of Cardiology, increased at-home BP monitoring is essential for recognizing inconsistencies in measurements taken in a medical setting [4]. Currently, the predominant device for measuring BP is a mercury sphygmomanometer which involves attaching an inflatable cuff around the upper arm [5]. This process requires significant user effort, which limits the frequency of BP measurements and increases the chance of measurement error. The use of an arterial catheter can provide continuous BP measurement; however, it is highly invasive and impractical for daily life. On the other hand, wearable devices are widely used for non-invasive, continuous monitoring of biological information [6]. Continuous and automated blood pressure estimation could be incorporated into one's daily routine to obtain better insight and detect abnormal BP fluctuation.

One prominent approach is to estimate BP with the photoplethysmogram (PPG) sensor, which is available in most wrist wearables. The principle of the PPG sensor is to optically measure the dilation and constriction of blood vessels. The resulting PPG signal is a fusion of heart activity, vascular relaxation processes, and microcirculation system status, making its time-frequency domain information rich and diverse [7]. In this paper, we propose a deep learning approach to personalized BP estimation based on the PPG signal.

Traditional machine learning approaches to PPG-based BP estimation focus on pulse wave analysis (PWA) methods. PWA involves extracting both time and frequency domain features from the PPG series and using these hand-crafted features as inputs to the BP estimation model. [8] extracts nineteen features from each PPG cycle based on its morphology. They use these features and the corresponding SBP and DBP values to train different regression models. Their approach lacks personalization, which may be the reason for higher estimation errors since these features have a person-specific response to BP [9]. [10] and [11] both use a random forest as their BP estimation model. [10] uses a feature selection algorithm to determine which morphological features are most useful for BP estimation and found that many features are irrelevant. Since the PPG signal is highly sensitive to different sources of noise [12] and its morphology can range from person to person, it is difficult to detect the key points in the signal required for feature engineering. In addition, manually engineered features can prove to be redundant or irrelevant in the PPG-BP modeling process. As a result, the information contained in the PPG signal may not be fully utilized.

In our previous work [13], we propose a method for personalized BP estimation using wavelet decomposition to extract time-frequency domain features from the PPG signal. These features are then used to train a random forest model for SBP and DBP estimation. Unlike previous

approaches which extract features from the PPG signal on a per cycle basis, wavelet decomposition captures dependencies between cycles in the time-frequency domain. While this approach produced accurate estimations, 10 hours of continuous BP and PPG data are required per person for training. Although PPG data can be continuously measured, large amounts of BP data are difficult to acquire outside a hospital setting.

In order to address the limitations of these previous methods, we propose a deep learning approach that utilizes a novel transfer learning technique that requires as little as 50 samples to train accurate personalized models. Deep learning models are widely used to model nonlinear relationships and have been applied to various tasks involving physiological signals [14-16]. Deep learning addresses the challenges of manual feature engineering and information loss by directly learning from the raw PPG data. [17-19] build deep learning models for PPG-based BP estimation and utilize personalization techniques to improve performance. [17] uses a spectro-temporal neural network that takes a 5 second PPG segment and its corresponding spectrogram as inputs to their model. When personalizing their model, the SBP and DBP MAE decrease by 39% and 44%, respectively, indicating that the relationship between BP and PPG is subject-dependent. [18] utilizes a Siamese neural network to estimate the offset from a calibration PPG-BP sample. The network uses a series of convolutional layers to derive an effective representation of the PPG series and achieves high estimation performance. [19] proposes a convolutional neural network (CNN) for BP estimation and utilizes transfer learning to personalize their model to each patient. Their proposed model requires 4000 personal BP samples for transfer learning to achieve high performance. Such a large number of personal BP samples is not possible to collect outside a hospital setting.

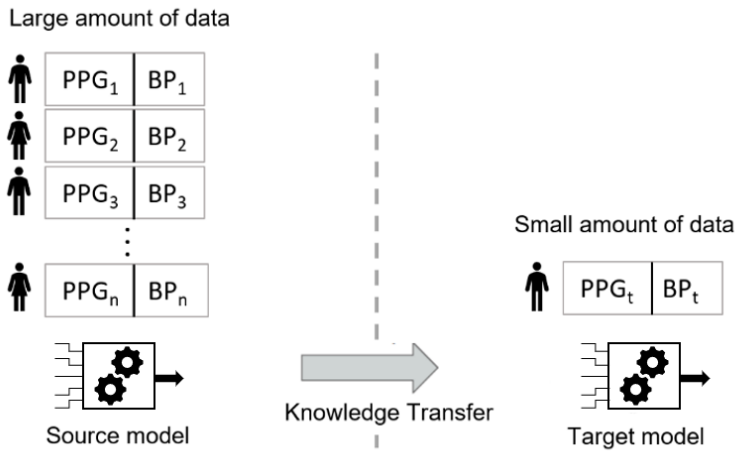

Figure 1.1 Transfer learning overview for PPG-based BP estimation.

Transfer learning focuses on storing knowledge gained from solving one problem (i.e., source domain) and applying it to a different but related problem (i.e., target domain), which usually contains a small number of data samples to train a model [20]. We propose to use a pre-trained model with abundant PPG and BP data from a large pool of source patients to drastically reduce the required data for new patients, as illustrated in Figure 1.1.

Deep learning models are conducive to transfer learning due to the modularity of their architectures. In this work, we develop our architecture, namely Blood Pressure – Convolutional Recurrent Neural Network (BP-CRNN), based on the Convolutional, Long Short-Term Memory, fully connected Deep Neural Network (CLDNN) [21], one of the popular hybrid artificial neural network (ANN) architectures. Our proposed method, namely BP-CRNN-Transfer, personalizes specific network layers during transfer learning to reduce the number of required training samples. Our contributions are as follows:

- We propose a hybrid neural network consisting of convolutional and recurrent layers which operate directly on the raw PPG time series to reduce information loss and effectively model the PPG-BP relationship.

- We propose a novel transfer learning technique that personalizes specific layers of a pre-trained network to improve the performance of PPG-based BP estimation, demonstrating that PPG-BP data of other patients can be used to enhance the modeling of a new patient's PPG-BP relationship.
- We demonstrate that the proposed transfer learning technique improves BP estimation performance by 23.3% for SBP and 19.1% for DBP. We verify our approach is consistent with the gold-standard BP measurement method through Bland-Altman and correlation analysis.
- We show that our proposed transfer learning method requires 10x less personal PPG-BP data to achieve performance equivalent to that of a new personalized model trained with abundant data.

The rest of the paper is organized as follows. In Section II, data acquisition and our network architecture are presented. We then detail the proposed transfer learning technique. In Section III, the performance of the proposed method is evaluated. We compare how model performance changes for different numbers of training samples, with and without using transfer learning. Finally, we conclude the paper in Section IV.

1.2 Method

In this section, we first describe the MIMIC III Matched Subset database and the PPG and BP preprocessing steps. We then present the network architecture and transfer learning technique.

1.2.1 Data Acquisition and Preprocessing

Data was obtained from the Multiparameter Intelligent Monitoring in Intensive Care III (MIMIC III) Matched Subset database [22,23]. This database contains records for thousands of intensive care unit patients. Records in this database have been matched to records from the

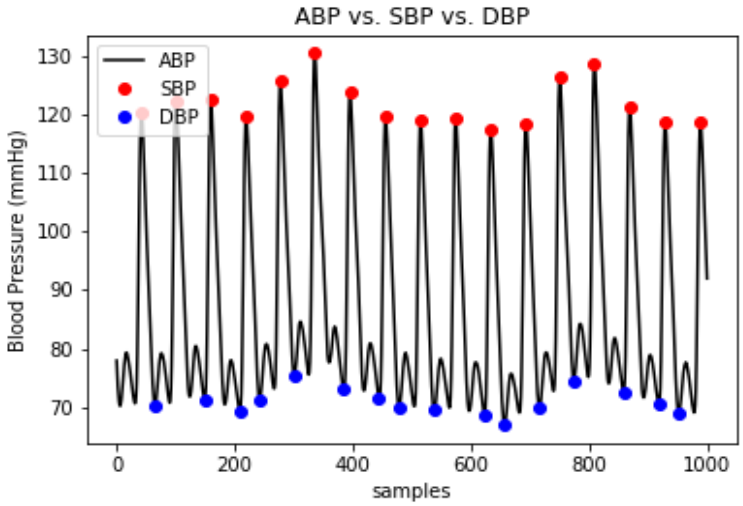

Figure 1.2 Output of peak detection algorithm – SBP and DBP vs. raw ABP time series.

MIMIC III Clinical database [24], which includes de-identified demographic data. The waveforms collected include ECG, respiration, continuous blood pressure, and PPG signals each sampled at 125 Hz. The arterial blood pressure (ABP) was directly measured from a radial artery using an invasive catheter. A fingertip sensor was used to measure the PPG data. Only patients with sufficient PPG and blood pressure data were considered for this study. We trained and tested our PPG-based BP estimation method on 100 randomly selected patients who had at least 10 hours of high-quality data after preprocessing. Out of these 100 patients, 56 are male and 44 are female. The age of the patients ranges from 21 to 82 with a mean age of 58.

Our objective is to operate directly on the raw PPG data and estimate SBP and DBP simultaneously. The first stage of data preprocessing involves splitting the raw PPG signal into 5-second segments and down sampling from 125 Hz to 25 Hz as this covers the important frequency components [25]. Next, each PPG segment is labeled with the mean SBP and DBP during that segment. SBP and DBP values are obtained from the raw ABP series using a peak detection algorithm as illustrated in Figure 1.2. Figure 1.3 describes the distribution of SBP and DBP samples. Some sections of the PPG series are corrupted due to motion artefacts or because the

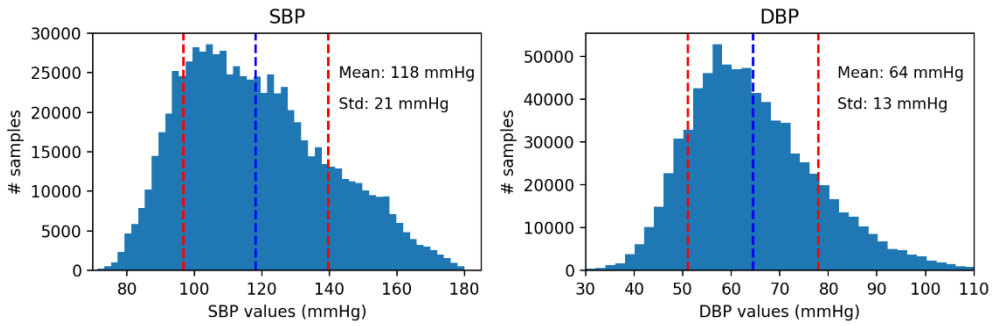

Figure 1.3 Distribution of SBP and DBP samples among the 100 patients. The blue dashed lines indicate the mean SBP/DBP and the red dashed lines correspond to 1 standard deviation above and below the mean SBP/DBP.

patient was not properly wearing the sensor. In order to discard these corrupted sections, an autocorrelation filter is implemented. Since an uncorrupted PPG segment should maintain a high degree of periodicity, it is expected that the signal's autocorrelation is high when the segment is offset by multiples of the cycle length. Figure 1.4 displays both an uncorrupted and corrupted PPG segment and the corresponding autocorrelation signals. The peaks in the autocorrelation signal are used to determine the quality of each PPG segment. An empirical threshold of 0.7 was set on the maximum autocorrelation. The filtered PPG segments are then normalized to zero mean and unit variance. Using this labeled dataset, we train our proposed personalized deep neural networks for BP estimation.

1.2.2 Network Architecture

We propose a hybrid network architecture, namely BP-CRNN, that makes use of convolutional layers, a gated recurrent unit (GRU), and fully connected (FC) layers. This is an adaptation of the CLDNN network presented in [21]. Instead of a LSTM, we use a GRU which behaves nearly identically with one fewer equation. In addition, we pass the outputs of both the first and third convolutional layers to the GRU. Figure 1.5 displays our architecture. The rationale is as follows: The convolutional layers serve as feature extractors for the raw PPG input, while the

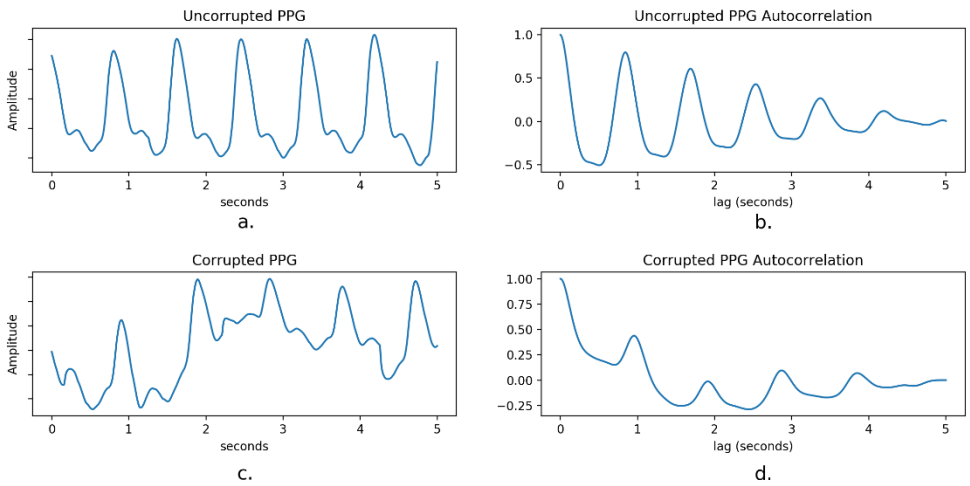

Figure 1.4 Comparison of (a) an uncorrupted PPG segment and (b) its corresponding autocorrelation signal to (c) a corrupted PPG segment and (d) its corresponding autocorrelation signal.

GRU models the temporal dependencies between these features. The GRU's outputs are then fed to the fully connected layers which transform the features into a space that makes the BP easier to estimate.

The input PPG segment is convolved with 50 different filters to generate 50 outputs in the temporal-feature domain. The following two convolutional layers also contain 50 filters, which are convolved with these features to generate the final features from the PPG segment. Each layer is followed by a rectified linear unit (ReLU) activation function. The output feature maps of each convolutional layer are calculated using the equation:

$$x_j^l = Relu\left(\left(\sum_i x_i^{l-1} * k_{ij}\right) + b_j^i\right) \quad (1)$$

where x_j^l is the j_{th} map generated by the convolutional layer l, x_i^{l-1} is the i_{th} feature map of the previous convolutional layer $l-1$, k_{ij} represents the i_{th} trained convolution kernel, b_j^i is the additive bias, while $*$ represents the convolution operation and *Relu* is the activation function.

Stacking convolutional layers results in a learned feature hierarchy, where initial layers extract lower-level features and deeper layers extract higher-level features [26]. We varied the

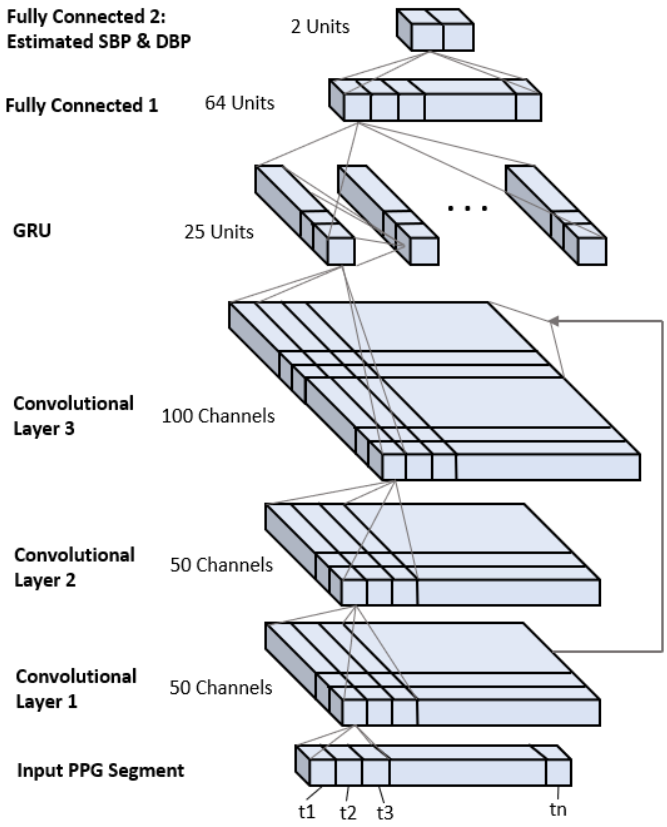

Figure 1.5 Proposed BP-CRNN architecture– Convolutional layers serve as feature extractors, GRU models temporal relationship between features, and fully connected layers transform GRU outputs to SBP and DBP.

number of convolutional layers from 1 to 5 and found that 3 convolutional layers resulted in the best performance. In order to provide both low and high-level features to the GRU to process simultaneously, the outputs of the first and third convolutional layers are concatenated. Since each convolutional layer contains 50 filters, 100 extracted feature series are passed to the GRU. The extracted features at each level are padded such that they have the same length as the input PPG sequence. As a result, the input to the GRU has a shape of $100 * t_n$ where t_n is the length of the input PPG segment. The GRU is able to learn the temporal relationship between these multiple feature channels. A GRU consists of gating units that control the flow of information within the module [27]. The following equations describe the operation of the GRU:

$$z_t = \sigma\left(W^{(z)}x_t + U^{(z)}h_{t-1}\right) \tag{2}$$

$$r_t = \sigma\left(W^{(r)}x_t + U^{(r)}h_{t-1}\right) \quad (3)$$

$$h_t^{'} = tanh\left(W^{(h)}x_t + U^{(h)}(r_t \odot h_{t-1})\right) \quad (4)$$

$$h_t = z_t \odot h_{t-1} + (1 - z_t) \odot h_t^{'} \quad (5)$$

In Eq. (5), the final GRU activation h_t is a linear interpolation between the previous activation h_{t-1} and candidate activation h_t' where the update gate z_t determines how much the unit updates its activation. $\odot$ represents element-wise multiplication. Eq. (2) describes the update gate z_t calculation, where $W^{(z)}$ and $U^{(z)}$ are each a set of trainable weights that process the input x_t and the previous activation h_{t-1}, respectively. σ represents the sigmoid function. The candidate activation h_t' is calculated in Eq. (4), where r_t represents the reset gate, $W^{(h)}$ and $U^{(h)}$ represent trainable sets of weights, and $tanh$ represents the hyperbolic tangent function. When r_t is close to 0, the reset gate enables the unit to forget the previous activation h_{t-1} when calculating the candidate activation h_t' [27]. In Eq. (3), the reset gate r_t is calculated similarly to the update gate. $W^{(r)}$ and $U^{(r)}$ represent the reset gate's trainable weights that process the input x_t and the previous activation h_{t-1}, respectively. At each time step, a 100-element vector is processed by the GRU, where each element corresponds to a feature value. A GRU activation size of 25 was experimentally determined to produce high performance, resulting in an output of shape $25 * t_n$.

The last two network layers are fully connected layers that carry out the final BP estimation. FC layers are effective at mapping features into a more separable space [26]. The activations of the GRU at each time step are flattened into a single vector for the first FC layer to the process. The output of the network is a 2-dimensional vector corresponding to the estimated SBP and DBP values. A ReLU activation function is again used after each FC layer. Batch normalization [28] is utilized to stabilize the input distribution of each layer during training. This reduces internal covariate shifts and results in faster training. Overall, this architecture realizes the high level of complementarity these individual neural network layers exhibit.

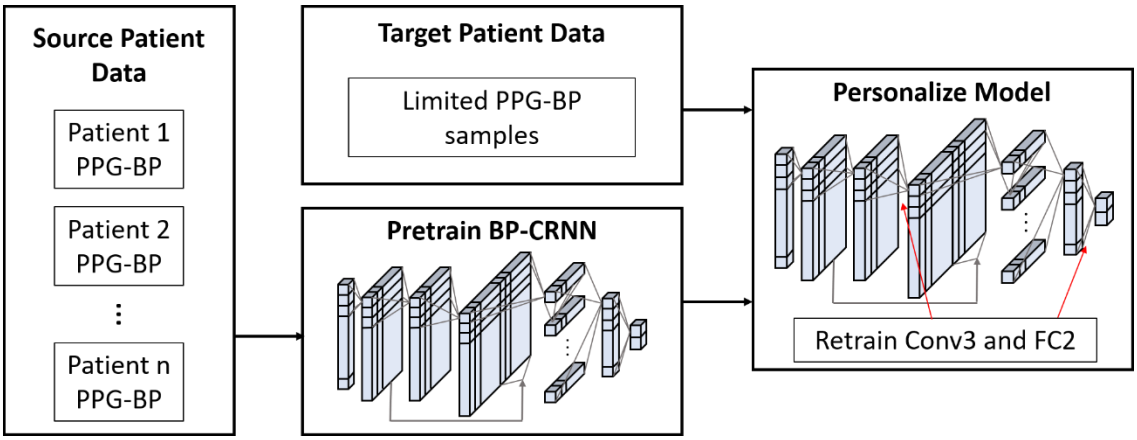

Figure 1.6 Proposed transfer learning method, namely BP-CRNN-Transfer. A BP-CRNN model is first pretrained using abundant source patient data. The final convolutional layer and fully connected layer are finetuned with the target patient's data.

1.2.3 Transfer Learning

To train deep neural networks, a large amount of training data is required to learn effective feature representations. Since our goal is to train personalized PPG-based BP estimation models, this means many data samples from a single individual are required. While PPG data can continuously be collected via a noninvasive wearable, BP data is more difficult to collect. In order to address this, we propose a transfer learning technique that results in high performance even when limited data from the target patient is available.

Transfer learning has most notably been applied to computer vision (CV) and natural language processing (NLP) tasks. [29] argues that physiological signals share two important commonalities with CV and NLP: consistency and complexity. Physiological patterns are consistent across individuals but complex enough that learning effective feature representations is nontrivial. [30] describes how initial convolutional layers extract lower-level features, which can be shared across tasks, while deeper layers generate higher-level features which are task-specific. In addition, training with different tasks (patients in our case), can result in a more powerful representation of the data that could not be learned from a single task (patient). Inspired by [29,30],

we first train our model with PPG-BP data from a variety of individuals to learn robust feature extractors that can be transferred between patients.

Figure 1.6 illustrates our proposed transfer learning process, namely BP-CRNN-Transfer. PPG and BP data from n source patients is used to pre-train a BP-CRNN model. This network is then used as an initialization for finetuning. In order to personalize the model, data from the target patient is used to finetune specific layers in the network. The last convolutional layer (Conv3) and last fully connected layer (FC2) are retrained using the target patient's data. In addition, the batch normalization parameters are updated to account for the different data distribution of the target patient. It was experimentally determined that retraining these two specific layers resulted in the most robust transfer learning performance. Table III in Sec. III (B) describes the transfer learning performance for different combinations of personalized layers. By retraining the final convolutional layer, the network can learn high-level PPG feature representations specific to the individual. Finetuning the last FC layer allows the model to learn the relationship between the extracted features and BP for the patient of interest. Our BP-CRNN model consists of approximately 250,000 trainable parameters, where 18,000 of these parameters are within the two layers we finetune. This indicates that we only need to update 7.2% of the network parameters learned from the source dataset. By retraining a small percentage of parameters, we prevent the network from overfitting to the limited target training data.

1.3 Results and Discussion

In this section, we describe the experiment settings and compare our personalized BP estimation results with and without transfer learning to previous methods. We examine how performance is affected by the number of personal data samples used during training and demonstrate that our transfer learning approach can achieve high performance with limited data.

We verify our approach is consistent with the gold-standard BP measurement method through Bland-Altman and correlation analysis.

1.3.1 Experiment Setting

We implement and evaluate our deep learning model using the Pytorch library [31] in the python environment on an Intel i5 3.2GHz quad-core and 16GB RAM computer. Nvidia GeForce GPUs are utilized to carry out network training. 1-dimensional filters of size 7 were implemented for each convolutional layer and zero padding was used to maintain the input PPG dimension. Based on the results from [32], a large range in the number of filters will result in similar performance before overfitting occurs. We chose to use 50 filters at each layer. All networks are trained using the Adam optimizer [33]. 10 hours of PPG and BP data are selected from each patient to be used in our experiments. 5-fold cross-validation is carried out for each patient separately. This involves shuffling each patient's data and using 5 different train, validation, and test splits for each experiment. Each validation and test set comprises of 1 hour of PPG-BP data. The number of samples included in the training sets is varied from 50 to 3600 samples in order to assess how performance is affected by training set size, which is detailed in Sec. III (C). Data separation between patients is maintained to ensure that no personal data from the target patient is used in pretraining for transfer learning. Mean absolute error (MAE) is calculated and used as our evaluation metric. For each experiment, we provide the average of MAEs over all patients. MAE is defined as follows:

$$MAE = \frac{\sum_{i=1}^{n} \left| BP_{pred}^{i} - BP_{actual}^{i} \right|}{n} \tag{6}$$

For our non-transfer method, namely BP-CRNN, separate personalized models are trained for each of the 100 patients. Each model is trained only using data from the individual patient. Since we do not use transfer learning, the parameters of the initial model are randomly initialized

and all layers are updated during training. To train these models, we use 0.01 as the learning rate and 32 as the batch size.

For testing our transfer learning technique, namely BP-CRNN-Transfer, the initial model for the first 50 patients is trained with the data of the last 50 patients, and vice versa. This ensures that no data from the target patient is used during pretraining. When training the initial model for transfer learning, the learning rate and batch size are set to 0.001 and 256, respectively. In this case, the learning rate can be decreased and the batch size increased because there is much more training data, resulting in a greater number of update steps per epoch. When fine-tuning the pre-trained model to the target patient, the learning rate and batch size are set back to 0.01 and 32, respectively, and only the specific layers mentioned in Sec. II (C) are updated. Early stopping [34] is implemented for every training session to save the learned network weights once the error on the validation set begins to increase. Each network is trained 5 times and the results averaged in order to account for differences in model convergence. Our model's inference time is 0.32 ± 0.09 (mean ± std) seconds. This time is based on implementation on a Nvidia GPU. In our future work, we plan to investigate a lightweight model that can be directly implemented on a wearable device and research the tradeoffs between model accuracy, inference time, and memory requirements.

1.3.2 BP Estimation Results

We compare the BP estimation performance of our personalized models without and with transfer learning to that of an aggregate model and previous methods in Table 1.1. BP-CRNN and BP-CRNN-Transfer correspond to our personalized approach without and with transfer learning, respectively. The aggregate model, namely Aggregate BP-CRNN, is trained in the same fashion as the pre-trained models for transfer learning as described in the previous section. However, no personalization or transfer learning is applied. The high estimation error of Aggregate BP-CRNN

Table 1.1 Comparison of BP estimation methods.

Method	SBP MAE (mmHg)	DBP MAE (mmHg)
Aggregate BP-CRNN	16.3	8.46
Mean Regressor	9.07	4.58
RF - Wavelet [13]	4.88	2.61
Spectro-Temporal NN [17]	9.43	6.88
Siamese NN [18]	5.95	3.41
CNN-Transfer [19]	4.06	2.20
BP-CRNN	4.59	2.72
BP-CRNN-Transfer	**3.52**	**2.20**

demonstrates the requirement for personalization in order to effectively model the PPG-BP relationship.

Next, we compare our proposed approach against a dummy regressor, namely Mean Regressor, which always predicts the mean SBP and DBP from the target patient's training set. This is an important comparison to make as there may be a subject with relatively constant BP, in which case the BP-CRNN's estimation errors will be low [17]. This comparison is drawn to ensure that our model has learned more than simply estimating the patient's mean BP. In addition, we compare our approach to our previous work and to the latest deep learning approaches that propose personalized BP estimation methods. In our previous work, we apply wavelet decomposition to the PPG series for feature engineering and train a random forest (RF) as our BP estimation model [13]. As mentioned in the introduction section, [17] trains a spectro-temporal neural network using personal data samples from each patient. [18] uses a Siamese neural network that takes a raw PPG segment as input and estimates the BP offset from a calibration PPG-BP sample. [19] trains a convolutional neural network for BP estimation and utilizes transfer learning to personalize their model to each patient.

In our current approach, a model is trained for each patient using both a non-transfer learning and transfer learning approach, as described in the experiment setting. Without transfer learning, namely BP-CRNN, we achieve an average MAE of 4.59 and 2.72 mmHg for SBP and DBP, respectively. As shown in Table I, even without using transfer learning, our proposed model achieves improvement in SBP performance compared to the methods presented in [13,17,18]. We attribute this improvement to the complementarity of our network architecture and its ability to reduce information loss by operating directly on the raw PPG series. With the transfer learning approach, namely BP-CRNN-Transfer, the MAEs decrease to 3.52 and 2.20 mmHg corresponding to a 23.3% and 19.1% increase in performance for SBP and DBP estimation as compared to our non-transfer method. The performance achieved by our BP-CRNN-Transfer method is also better than our previous approach RF-wavelet [13] as well as previous deep learning methods [17-19]. We achieve a 27.9% and 15.7% improvement from [13], 62.7% and 68% improvement from [17], and 40.8% and 35.5% improvement from [18] for SBP and DBP, respectively. We achieve a 13.3% improvement for SBP MAE and the same DBP MAE as compared to [19]. We attribute this increase in performance to the specific layers we finetune during transfer learning and our network's ability to effectively store information contained in source patients' data. The BP-CRNN-Transfer MAE is well under the Mean Regressor MAE, which is 9.07 mmHg for SBP and 4.58 mmHg for DBP, indicating that the model can learn a meaningful relationship between PPG and BP. Since [19] achieves the closest performance to our proposed method, we reimplement their approach in order to perform statistical tests. We carry out a Paired Student's t-Test separately for each patient to assess the statistical significance of the difference in estimation errors between our method and [19]. For 84 out of the 100 patients, the difference in performance is statistically significant at the level 0.05 for both SBP and DBP.

Table 1.2 Comparison of proposed method to BHS Standards. Both our non-transfer (BP-CRNN) and transfer learning (BP-CRNN-Transfer) approaches achieve Grade A performance for SBP and DBP.

	SBP				DBP			
Method	≤ 5 mmHg	≤ 10 mmHg	≤ 15 mmHg	Grade	≤ 5 mmHg	≤ 10 mmHg	≤ 15 mmHg	Grade
BP-CRNN	72%	92%	97%	A	89%	98%	99%	A
BP-CRNN-Transfer	80%	95%	98%	A	93%	99%	100%	A

We evaluate our proposed method according to the British Hypertension Society (BHS) and the Association for the Advancement of Medical Instrumentation (AAMI) standards for BP measurement. The BHS standard assigns a performance grade based on the percentage of estimated BP samples that fall within 5, 10, and 15 mmHg of the corresponding reference BPs. To achieve Grade A accuracy, at least 60/85/95% of the estimated BP samples must have an absolute difference of ≤ 5/10/15 mmHg from the reference BPs, respectively [35]. Table 1.2 describes the results of our non-transfer and transfer learning approaches according to the BHS standards. For our non-transfer approach, 72/92/97% of estimated SBP samples have an absolute difference ≤ 5/10/15 mmHg, respectively. When using our transfer learning approach, these percentages increase to 80/95/98% of estimated SBP samples. For our non-transfer approach, 89/98/99% of estimated DBP samples have an absolute difference ≤ 5/10/15 mmHg, respectively. When using our transfer learning approach, these percentages increase to 93/99/100% of estimated DBP samples. Both approaches achieve Grade A performance according to the BHS standard for SBP and DBP.

The AAMI standard for accurate BP measurement requires that the mean error between estimated and reference BPs is ≤ 5 mmHg and the standard deviation (SD) of errors is ≤ 8 mmHg [36]. Figure 1.7 displays the error distribution for SBP and DBP using both our non-transfer and

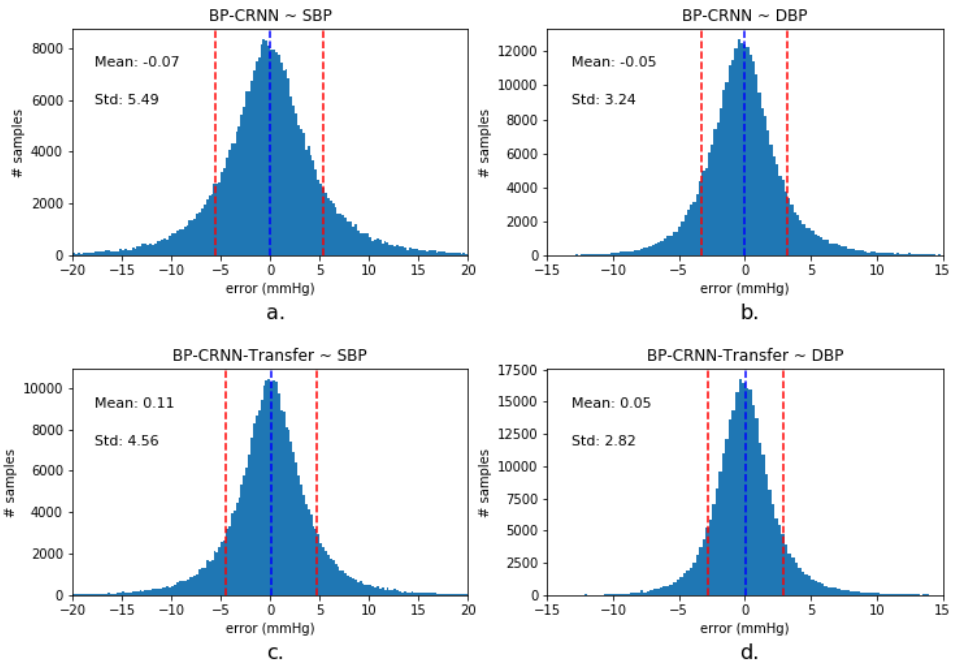

Figure 1.7 Distributions of (a) SBP and (b) DBP errors using our non-transfer approach compared to distributions of (c) SBP and (d) DBP errors using our transfer learning approach. The blue dashed lines indicate the mean error and the red dashed lines correspond to 1 standard deviation above and below the mean error.

transfer learning approach over all patients. Our BP-CRNN (non-transfer) approach achieves a mean error and standard deviation of -0.07 ± 5.49 mmHg and -0.05 ± 3.24 mmHg for SBP and DBP, respectively. Our BP-CRNN-Transfer approach achieves a mean error and standard deviation of 0.11 ± 4.56 mmHg and 0.05 ± 2.82 mmHg for SBP and DBP, respectively. The mean error for each approach is approximately 0 mmHg. When using our transfer learning approach, the SD of errors decreases from 5.49 to 4.56 mmHg and 3.24 to 2.82 mmHg for SBP and DBP, respectively. While both approaches satisfy the AAMI standard, our transfer learning approach achieves the requirement by a larger margin.

Table 1.3 compares the transfer learning performance when different sets of network layers are finetuned using target patient data. We use the first 10 patients in our dataset as target patients for this experiment. The source model is pre-trained with the last 50 patients' data. These results

Table 1.3 Comparison of transfer learning performance when finetuning different network layers.

BP-CRNN Layers Personalized	SBP MAE (mmHg)	DBP MAE (mmHg)
FC1, FC2	5.16	2.87
FC2	4.41	2.63
Conv1, Conv2, Conv3, GRU, FC1, FC2	4.37	2.41
Conv3, FC1, FC2	4.32	2.46
Conv2, Conv3, GRU, FC1, FC2	4.28	2.38
Conv3, GRU, FC1, FC2	4.25	2.37
Conv3, GRU, FC2	3.90	2.28
Conv3, FC2	**3.84**	**2.24**

are averaged over the 10 target patients. Evidently, retraining only the final convolution layer (Conv3) and fully connected layer (FC2) results in the best transfer learning performance. If the Conv3 layer is not personalized, the SBP MAE increases from 3.84 to 4.41 mmHg and the DBP MAE increases from 2.24 to 2.63 mmHg. This demonstrates the importance of personalizing the last convolutional layer in order to learn higher level features specific to the individual. One interesting observation is that, on average, it is better not to retrain the GRU with the target data. The average SBP and DBP MAEs when finetuning the GRU layer with the Conv3 and FC2 layer are 3.90 and 2.28 mmHg, respectively. If the GRU is not personalized, the average SBP and DBP MAEs are 3.84 and 2.24 mmHg, respectively. This may be because the GRU is modeling the temporal relationship between features, and not the features themselves. This indicates that the temporal modeling of PPG features is transferable across individuals in addition to the lower-level convolutional filters.

Table 1.4 compares the transfer learning performance when different numbers of source patients are used for pretraining the initial model. Like the previous experiment, we use the first 10 patients in our dataset as target patients for this experiment and the results are averaged over

Table 1.4 Comparison of transfer learning performance when pretraining with different number of source patients.

# Source Patients	SBP MAE (mmHg)	DBP MAE (mmHg)
10	4.07	2.36
30	3.96	2.28
50	**3.84**	**2.24**
70	3.85	2.24
90	3.85	2.23

these patients. We compare the transfer performance when using 10, 30, 50, 70, and 90 source patients for pretraining. We finetune the "Conv3, FC2" layer set during the transfer learning step. We observe that the MAEs for SBP and DBP decrease as more source patients are included but level off at 50 patients. The MAEs for SBP estimation when using 50, 70, and 90 source patients are 3.84, 3.85, and 3.85 mmHg, respectively. The MAEs for DBP estimation when using 50, 70, and 90 source patients are 2.24, 2.24, and 2.23 mmHg, respectively. These results demonstrate that including more than 50 source patients does not enhance the transfer learning performance. This indicates that there is sufficient variability and information among 50 patients to learn effective transferable features for PPG-BP estimation.

1.3.3 Effect of Training Set Size

Next, we discuss how our non-transfer and transfer learning performances change based on the number of target patient training samples. We test the model performance using 5 different amounts of personal training data: 3600, 1800, 360, 100, and 50 data samples. Since each input PPG segment is 5 seconds, 3600 samples correspond to 5 hours of data. For each case, the validation and test sets are kept the same in order to ensure a fair comparison. Figure 1.8 displays the relationship between the number of training samples and SBP (left) and DBP (right) estimation performance. The blue curves correspond to our non-transfer approach, namely BP-CRNN, while

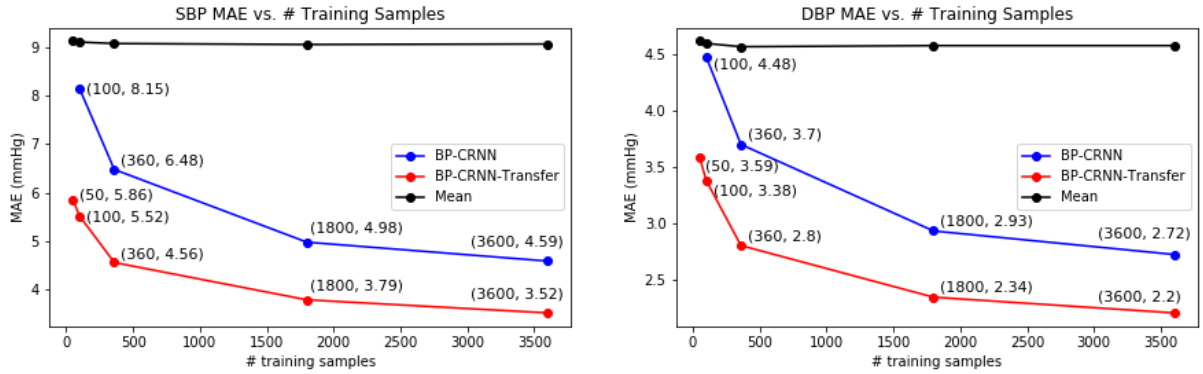

Figure 1.8 BP estimation performance for different training set sizes. The labeled points for 360 and 3600 training samples indicate that our BP-CRNN-Transfer method can achieve equivalent performance to the non-transfer BP-CRNN method with 10x less data.

the red curves correspond to our transfer method, namely BP-CRNN-Transfer. Each point is labeled with the number of training samples and corresponding MAE. The black lines represent the performance of the dummy Mean Regressor, which always predicts the mean SBP and DBP of the target patient's training set. Again, we use the Mean Regressor's performance as a reference to ensure our model is learning more than simply estimating with the patient's mean SBP and DBP.

Evidently, using transfer learning improves performance for each number of training samples. As the number of training samples is reduced, the MAE increases for both approaches, but at a lower rate when utilizing transfer learning. When training with 100 data samples using the non-transfer approach, the MAE increases to 8.15 mmHg for SBP and 4.48 mmHg for DBP. In this case, the error is approaching that of the Mean Regressor, meaning the model has difficulty learning the PPG-BP relationship. If further reduced to 50 training samples, the model is unable to converge. This is why there is no point plotted for 50 samples when using our non-transfer approach. On the other hand, when using 100 training samples, the performance of our transfer learning approach for SBP and DBP is 5.52 and 3.38 mmHg, respectively. This corresponds to a 32.3% and 24.6% performance improvement for SBP and DBP estimation when using our transfer

learning technique. By comparing the non-transfer approach using 3600 samples to the transfer approach using 360 samples, we can see that the MAE is similar for SBP (4.59 vs. 4.56 mmHg) and DBP (2.72 vs. 2.80 mmHg) estimation. This indicates that 10x less personal PPG-BP data is required by our proposed transfer learning approach to achieve performance equivalent to that of a new personalized model trained with abundant data. For 50 training samples the model is able to converge using transfer learning, resulting in a MAE of 5.86 mmHg for SBP and 3.59 mmHg for DBP. The cuff-based standard is a MAE of ≤5 mmHg for both SBP and DBP [37]. Hence, our transfer learning technique satisfies this requirement for DBP and misses this requirement by 0.86 mmHg for SBP, when using 50 training samples. These results demonstrate that accurate personalized models can be trained even with limited personal PPG and BP data.

1.3.4 Bland-Altman and Correlation Analysis

Bland-Altman analysis is a technique for comparing a new measurement device or procedure to an approved method [38]. The goal is to assess the extent to which two methods designed to measure the same parameter are in agreement. Here, the two methods for BP measurement being compared include the invasive arterial catheter and our BP-CRNN-Transfer model. The difference in measurements between these two methods is plotted against the average measurement of the two devices. The difference between methods and mean of methods are calculated for each data sample using equations 7 and 8, respectively.

$$BP_{diff} = BP_{catheter} - BP_{BP-CRNN} \tag{7}$$

$$BP_{mean} = \frac{BP_{catheter} + BP_{BP-CRNN}}{2} \tag{8}$$

It is common to compute the 95% limits of agreement between measurement methods. These limits are defined as the average difference between measurement methods (blue dashed line in Figure 9) ± 1.96*standard deviation of the differences between measurement methods (red-

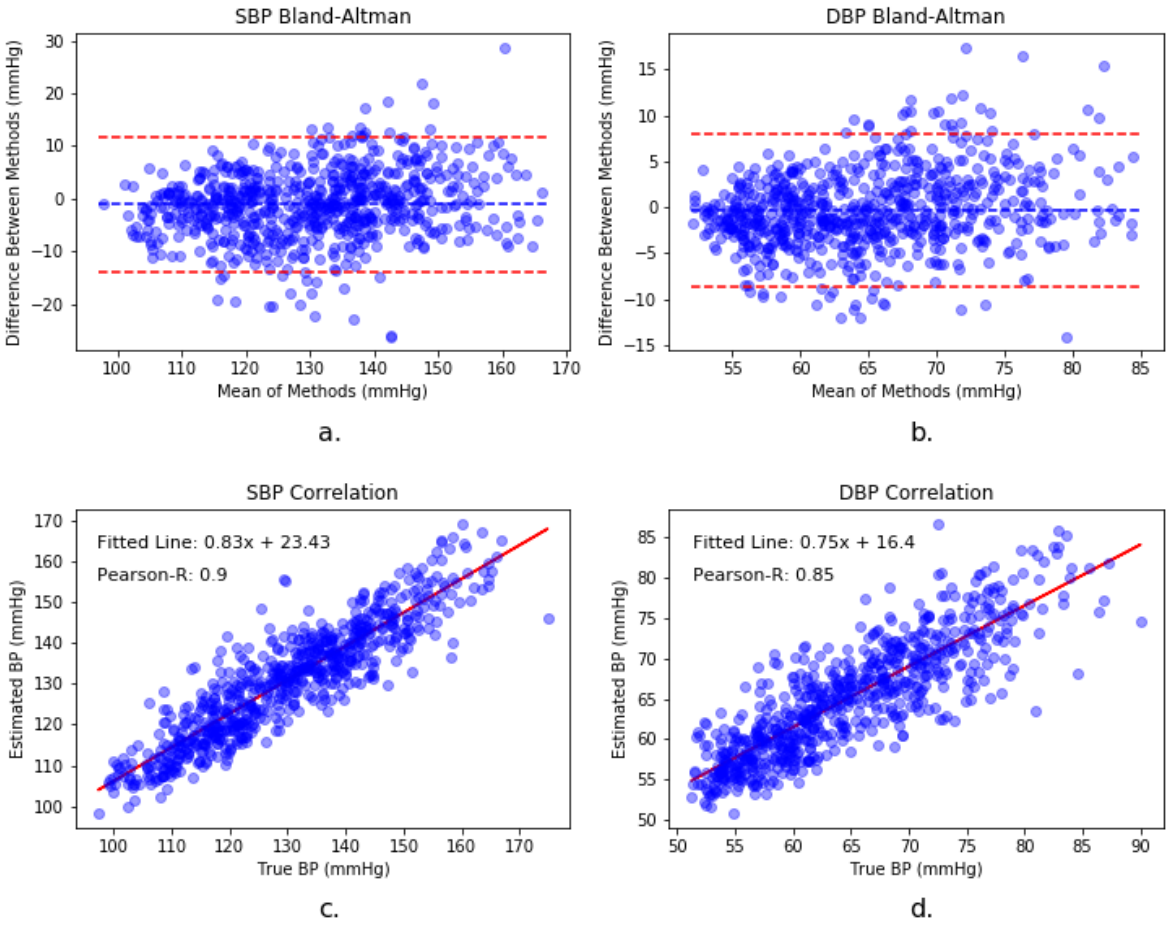

Figure 1.9 Bland-Altman and Pearson correlation analysis for one patient used to assess agreement between BP measurement methods. Plots (a) and (b) display Bland-Altman analysis for SBP and DBP, respectively. Plots (c) and (d) display the correlation between estimated and reference SBPs and DBPs, respectively.

dashed lines in Figure 1.9). For two methods to be considered comparable, Bland-Altman recommends that 95% of the samples should fall within these limits (red dashed lines). Among all 100 patients, 86% and 93% achieve this agreement for SBP and DBP measurement, respectively.

We also carry out Pearson correlation analysis [39] separately for each of the 100 patients to compare our method's estimated BP to the reference BP. The Pearson-R correlation coefficient is a measure of how linearly correlated two sets of data are. When using our non-transfer approach, the average and standard deviation of the Pearson-R coefficient is 0.83 ± 0.10 and 0.73 ± 0.17 for SBP and DBP, respectively. When using our transfer learning approach, the average and standard

deviation of the Pearson-R coefficient is 0.90 ± 0.06 and 0.82 ± 0.12 for SBP and DBP, respectively. This increase in correlation again shows the ability of transfer learning to improve estimation performance.

Since it is not possible to show individual plots for each patient, we provide plots for one patient whose Pearson correlation is similar to the average correlation across all patients. Figure 9 displays both the Bland-Altman and correlation plots for SBP and DBP for this patient. 95.1% of the SBP differences and 95.6% of the DBP differences fall within the Bland-Altman limits of agreement. The correlation between estimated and reference BPs is 0.9 and 0.85 for SBP and DBP, respectively. These results demonstrate a high level of agreement between our model's estimated BP and the invasively measured BP from the arterial catheter.

1.3.5 Investigating Source Patient Selection

In this section, we discuss findings regarding source patient selection for individual target patients. Table IV compares results when using different numbers of source patients, however, these results represent an average and do not capture performance variations at the individual patient level. The goal of this experiment is to determine whether there are optimal smaller sets of source patients for individual target patients.

In order to determine the effect of using different source patients for individual target patients, multiple models are pre-trained. Table V displays the results for 3 different target patients, using 3 different pre-trained models for transfer learning. Model 1 represents the same initial model used in the previous experiments pre-trained with 50 source patients. Models 2 and 3 were pre-trained using different random sets of 10 source patients. For this experiment, 50 samples from the target patient are used to finetune each model.

Table 1.5 Transfer performance of different pretrained models.

	SBP MAE / DBP MAE (mmHg)		
	Patient 1	Patient 2	Patient 3
Model 1	4.73 / 3.27	7.96 / 4.99	**4.79 / 2.47**
Model 2	**4.47 / 3.18**	7.06 / 4.58	5.46 / 3.29
Model 3	4.76 / 3.20	**6.85 / 4.41**	5.05 / 3.12

On average, pretraining with 50 source patients (shown in Table 1.4) is better than pretraining with 10 source patients. However, for individual target patients, there may be certain smaller sets of source patients that result in better transfer learning performance, as shown in Table 1.5. This performance increase can be significant, especially seen for Patient 2. Model 3 (pre-trained with 10 source patients) performs 13.9% and 11.6% better for SBP and DBP estimation compared to Model 1 (pre-trained with 50 source patients) for this target patient. These results indicate that transfer learning performance can be further improved by selecting a specific subset of source patients for individual target patients. In future work, we plan to investigate this idea of intelligent source patient selection for improving transfer learning performance.

1.4 Conclusion

In this chapter, we present an effective hybrid network architecture for personalized BP estimation using the PPG signal. In order to reduce the number of personal PPG-BP samples required for training, we provide a novel transfer learning approach that personalizes specific layers of the network. Our method is tested over a demographically diverse set of patients, and our estimation performance achieves the BHS and AAMI standards.

In this study, the training and inference are implemented on a personal computer. For future work, we will investigate a light-weight BP estimation model which can be implemented directly on a wearable device that collects PPG data while providing comparable performance to our

current work. This will provide more real-time measurements and address concerns regarding data transmission and data privacy. BP measurement based on the PPG signal will enable a deeper understanding of how BP changes throughout the day, allowing the user to make adjustments in order to reach and maintain a healthy BP.

In the next chapter, we discuss the limitations of current remote monitoring practices for COVID-19 patients, predominantly reliant on manual symptom reporting and patient compliance. We present a ML-based remote monitoring method to estimate patient recovery from COVID-19 symptoms using automatically collected wearable device data, instead of relying on manually collected symptom data.

Chapter 1, in part, is from the material as it appears in the IEEE Journal of Biomedical and Health Informatics, 2022, Leitner, Jared; Chiang, Po-Han; Dey, Sujit. The dissertation author was the primary investigator and author of this paper.

Chapter 2 Classification of Patient Recovery from COVID-19 Symptoms using Consumer Wearables and Machine Learning

2.1 Introduction

Around the world, healthcare systems have been overwhelmed by the high numbers of COVID-19 cases, which has surpassed 437 million as of March 2, 2022 according to the World Health Organization (WHO) [40]. In the US, there were approximately 4.5 million COVID-19 hospitalizations between August 1, 2020 and February 28, 2022, according to the Center for Disease Control and Prevention (CDC) [41]. While this is a daunting number of hospitalizations, there have been approximately 80 million cases in the US [42], meaning most cases involve ambulatory patients being treated from home. This is an unprecedented number of patients needing care in their home and many are not being monitored in any way by medical personnel.

In order to combat this pandemic and provide more optimal care at scale, hospitals are changing the way in which healthcare is delivered. At the center of this changing landscape is a shift towards remote, continuous, and automated delivery of healthcare. This shift can lead to significant improvement in and scalability of at-home patient care for COVID-19, while at the same time enabling significant savings in human and equipment resources. Current remote monitoring for COVID-19 patients relies on manual symptom reporting, which is highly dependent on patient compliance. In this study, we demonstrate that data automatically collected from wearable devices together with machine learning (ML)-assisted diagnosis can enhance the efficiency and increase the scalability of remote monitoring for COVID-19 patients.

Wearable devices are one of the enabling technologies making this shift in healthcare delivery possible [43-46]. Consumer wearables, such as Apple Watch, Fitbit, and Samsung Galaxy Watch, remotely collect a great amount of lifestyle and vitals data in high granularity and

continuity. There is great opportunity for ML to assist in remote monitoring due to the large amount of data that is collected. Since it is not possible for doctors to manually review all remotely collected data [47], ML has the potential to provide automated insights into the health status of patients and significantly increase the scalability of remote patient care. This is especially helpful during a pandemic, where in-person interaction and monitoring may pose risks to healthcare workers and other patients. In addition, ML-assisted monitoring can provide patients with insights regarding their own progression, helping to keep them engaged and informed about their health.

Current research on using wearables and machine learning to combat COVID-19 is primarily focused on early detection of infection. The authors in [51-55] have demonstrated that it is possible to detect deviations in health data before significant symptoms arise. Using Fitbit devices, the researchers in [51] found that 26 out of 32 (81%) infected patients in their cohort had alterations in their heart rate, number of daily steps, or time asleep before becoming symptomatic. The authors in [54] used respiration rate, heart rate, and heart rate variability data collected from their wearable devices and proposed a deep learning method to estimate infection before the onset of symptoms. Early detection will enable individuals to quarantine earlier, helping reduce the spread of the virus. These studies demonstrate that wearable device data can provide actionable insights into the conditions of patients.

In this research, we propose a novel approach to estimate patient recovery from COVID-19 symptoms using automatically collected device data and machine learning. We partnered with the UCSD Health and Neighborhood Healthcare COVID-19 telemedicine clinics in order to carry out this research. Our remote monitoring system utilizes a Garmin wearable and symptom tracker mobile app for data collection and fuses this data into an online report for clinicians to review. We propose a novel labelling logic for patient recovery from COVID-19 symptoms using the symptom

tracker data. The labelling logic was developed in collaboration with UCSD Health doctors and the details are defined in Sec. III (B). Using this data, we train a patient recovery classifier which uses wearable data to estimate whether a patient has recovered from COVID-19 symptoms. We evaluate our method according to leave-one-subject-out (LOSO) CV to replicate the clinically relevant use case scenario in which a newly infected patient will not have data for model training. We compare the performance of different ML models and find that Random Forest (RF) is the top performing model. We propose a RF-based personalization technique in order to improve model performance. This technique utilizes the RF's weighted bootstrap aggregation algorithm in order to tune the model to each patient. The details are presented in Sec. III (D). Finally, we conduct Shapley Value analysis to inspect which device features have the greatest impact on classification. This analysis provides an interpretation of what the model has learned, which is especially important for medical applications. Our contributions are as follows:

- We deploy a remote patient monitoring system in two COVID-19 telemedicine clinics. The system consists of a wearable device, symptom tracker mobile app, and online dashboard which collects and analyzes vitals, lifestyle, and symptoms data. The estimated recovery status of each patient using our ML approach is displayed on the dashboard for clinicians to review.
- We propose a patient recovery classifier which uses wearable data to estimate whether a patient has recovered from COVID-19 symptoms. This ML tool can provide doctors with automated insights into the recovery status of their infected patients and bypass the need for manual daily symptom tracking.

- We carry out LOSO CV to mirror the clinically relevant use-case scenario and propose a RF-based personalization technique that improves model performance by tuning the model to each patient via weighted bootstrap aggregation.

The rest of the paper is organized as follows. In Section II, we investigate related works that utilize machine learning for COVID-19 diagnosis. In Section III, our remote monitoring system and data acquisition are presented. We then detail the proposed labelling logic and RF-based personalization technique for patient recovery classification. In Section IV, the performance of our proposed ML method is evaluated. In addition, we carry out top feature analysis based on Shapley Values and provide a discussion on research challenges. Finally, we conclude the paper in Section V.

2.2 Related Work

In this section, we present related research which is grouped into the follow categories: COVID-19 symptom tracking, early diagnosis, and recovery detection.

2.2.1 COVID-19 Symptom Tracking

The researchers in [48] utilize a smartphone-based app to collect symptom data from patients. In the app, patients also recorded when they had tested either negative or positive for COVID-19 infection. They propose a logistic regression model that combines the reported symptoms in order to predict COVID-19 infection. A combination of loss of smell and taste, fatigue, persistent cough, and loss of appetite resulted in the best model, which achieved a sensitivity and specificity of 0.65 and 0.78, respectively. The authors in [49] also used a mobile app for collecting symptoms data and COVID-19 test results. They trained a logistic regression model to predict COVID-19 infection based on self-reported symptoms, and calculated the odds ratio for each symptom in order to understand which symptoms were the strongest predictors.

Chills, fever, loss of smell, nausea, and shortness of breath were the top five strongest predictors of COVID-19 infection. Participants in their cohort with a positive test result experienced 5.6 symptoms on average. In [50], the researchers trained a gradient-boosting machine to predict COVID-19 infection based on 8 features: cough, fever, sore throat, shortness of breath, headache, age, sex, and known contact with an individual confirmed to have COVID-19. Their approach achieved a sensitivity and specificity of 0.86 and 0.79, respectively. Fever and cough were the top 2 features with the greatest impact on the model's prediction. These past works demonstrate that self-reported symptoms can be effectively used to predict COVID-19 infection. However, these approaches rely on patient compliance with manual symptom tracking. In contrast, wearable devices can passively collect data that is relevant to COVID-19 infection. In addition, wearable data can be predictive of COVID-19 infection prior to symptom onset.

2.2.2 Early Diagnosis of COVID-19

The authors in [51] use data collected from wearable devices for the early detection of COVID-19 infection. They propose an anomaly detection technique based on two parameters: 1. Resting heart rate (RHR), 2. Heart rate over steps (HROS). HROS was calculated by dividing heart rate by steps data at each hourly interval. They report that significant deviations in these parameters relative to the individual baseline can indicate COVID-19 infection. They utilize Gaussian density estimation to classify anomalies in the dataset. Their results show that 63% of COVID-19 cases in their cohort could have been detected before symptom onset. The researchers in [52] also utilize deviations from RHR to classify a patient as infected. They propose a deterministic finite state machine which triggers an alert when a patient's overnight RHR increases above the median of previous overnight RHRs by an empirically determined threshold. Their system generated alerts for 80% of the infected individuals prior to symptoms, however, many of the alert-generating

events were not associated with COVID-19 and instead attributed to other events, such as poor sleep, stress, alcohol consumption, intense exercise, or travel. While these studies demonstrate that deviations in physiological and activity data measured by wearable devices can be used for early detection of COVID-19, they only utilize a subset of possible device features (RHR and steps) and do not investigate ML-based approaches which are well suited to handle larger feature sets. Furthermore, they do not investigate if wearable device data can be used to monitor patient recovery from COVID-19.

The researchers in [53] trained a logistic regression model to differentiate COVID-19 positive vs. negative cases in symptomatic individuals based on symptoms and wearable device data. Baseline device data was calculated as the median of the data from 21 to 7 days before the onset of symptoms. They show an increase in model performance when including device data (RHR, sleep duration and step count) in addition to symptoms data as part of the feature set. The authors in [54] trained a convolutional neural network to predict illness given health metrics for that day and the preceding 4 days. These metrics included the mean respiration rate (RR) during sleep, mean heart rate during sleep, the root mean square of successive differences (RMSSD) of the nocturnal RR series and the Shannon entropy of the nocturnal RR series. They organize each data sample into 5x4 matrix and resize each matrix into a 28x28 image as the input to the network. Their method achieved a sensitivity and specificity of 51% and 90%, respectively. In [55], the researchers presented a gradient-boosting model based on decision trees to detect COVID-19 infection. Their approach achieved a sensitivity and specificity of 71% and 67%, respectively, when only using device features as input to the model. They grouped the device features into activity, sleep, and heart rate categories, and found that activity related features had the greatest impact on the model's prediction, followed by sleep and heart rate-related features. These works

demonstrate the ability of ML models to learn meaningful relationships between wearable device features and the onset of COVID-19 infection.

2.2.3 Recovery Detection from COVID-19

The research presented in [48-55] focused on predicting COVID-19 infection using self-reported symptoms or wearable device data. In contrast to these works, the objective of our research is to estimate recovery from COVID-19 symptoms using wearable device data. The researchers in [56, 57] present different approaches to estimate recovery from COVID-19 infection based on symptoms and demographic data. The authors train a support vector machine [56] and decision tree classifier [57] to estimate patient recovery based on symptoms, demographic, and travel-related features. In [56], the authors found that most of the patients who could not recover experienced a fever, cough, and fatigue. In [57], the authors extended their model to predict the number of days needed to recover from infection. Their model predicted a minimum of 5 days and a maximum of 35 days for COVID-19 patients to recover. Both approaches presented in [56, 57] rely on symptoms data and do not investigate the use of wearable device data for patient recovery estimation. We did not find any previous research that investigates whether wearable device data can be used to estimate patient recovery from COVID-19. This aligns with the observations of the authors in [58] who provide a review on the rise of wearables during the COVID-19 pandemic. None of the works presented in their review are focused on estimating patient recovery from COVID-19 symptoms. This motivates us to develop our own labeling logic for patient recovery in direct consultation with UCSD Health COVID-19 telemedicine doctors. In addition, the dataset we collect consists of a rich feature set spanning activity, sleep, stress, heart rate and SpO2 data. Our paper provides novel insights into which lifestyle and physiological signals are associated with patient recovery from COVID-19 symptoms.

2.3 Method

In this section, we first detail our study cohort and the proposed remote patient monitoring and reporting system. We then present the ML task of patient recovery classification and discuss its application. Finally, we describe the data preprocessing, the RF model, and our proposed personalization technique.

2.3.1 Clinical Study Cohort and eCOVID System

Our IRB-approved clinical study (protocol #181405) was in collaboration with UC San Diego Health and Neighborhood Healthcare, with patient enrollment, onboarding and management conducted by the Altman Clinical & Translational Research Institute at UC San Diego. The study was conducted starting in May 2020. Patients who tested positive for COVID-19 at each location were referred to our study coordinator. Eligible patients were required to be over 18 years old and stable for monitoring in an ambulatory setting, as determined by healthcare personnel at the point of care when testing was initially ordered. The characteristics of the included cohort are shown in Table 2.1. Subjects digitally consented using our symptom tracker mobile app, and those who did were provided a Garmin Vivosmart4 wearable device [59] to collect their lifestyle and vitals data for the study duration of up to 3 months. One of the deciding factors in using this device for this study is its ability to measure blood oxygen saturation (SpO2). Based on the findings of [60] and our discussion with UCSD Health doctors, SpO2 is a critical metric in determining the condition

Table 2.1 Cohort Statistics (n = 30).

	UCSD Health	Neighborhood Health
Total	23	7
# Men	11	3
# Women	12	4
Age (years, mean ± SD)	44.5 ± 13.1	31.6 ± 13.5

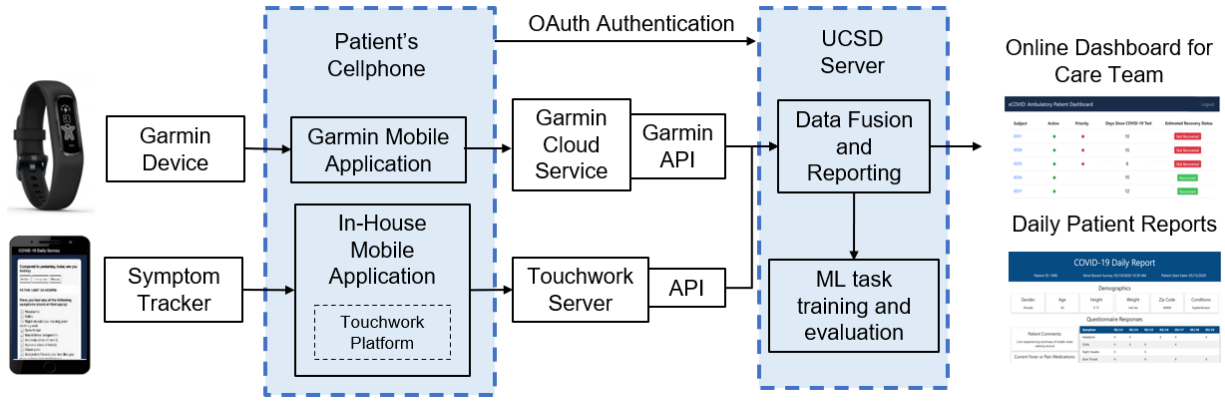

Figure 2.1 eCOVID remote monitoring and reporting system architecture.

of a COVID-19 infected patient. Figure 2.1 displays the overall architecture of our remote monitoring system, namely eCOVID. The system consists of a symptom tracker mobile app, developed using the Touchwork platform, and the Garmin device. The daily questions in the symptom tracker app were developed in collaboration with doctors at the UCSD Health COVID-19 telemedicine clinic and are detailed in Table 2.2. The vitals and lifestyle data collected by the Vivosmart4 wearable are detailed in Sec. III (C). Data was collected remotely through Garmin's application programming interface (API) [61].

Table 2.2 Daily Questions in Symptom Tracker App.

Questions *(Answers)*
1. How do you feel compared to yesterday? *(Better, Same, Worse)*
2. Have you had any of the following symptoms? *(Headache, Chills, Night sweats, Sore throat, Nasal/sinus congestion, Anosmia, Ageusia, Chest pain, Subjective fevers)*
3. How would you rate your fatigue? *(0-5)*
4. How would you rate your cough? *(0-5)*
5. How would you rate any shortness of breath? *(0-5)*
6. Are you able to drink & eat? *(Yes, Somewhat, Little, Minimal)*
7. What fever/pain medications have you taken?
8. What cough/breathing medications have you taken?

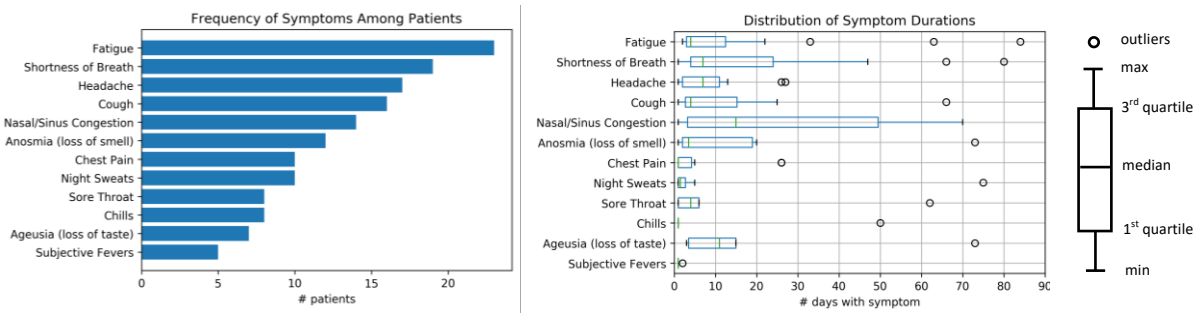

Figure 2.2 The left plot displays the number of patients who reported at least 1 day of the symptom. The right plot displays the distribution of the number of days each symptom was reported per patient. Only patients who reported the symptom are included in this distribution.

Figure 2.2 details the distribution of symptoms among patients and describes how long each symptom lasted. For fatigue, shortness of breath and cough, we marked the symptom as present if the patient reported a severity score of 2 or greater. The bar graph in Figure 2.2 displays the number of patients that experienced each symptom. Fatigue, shortness of breath and headache were the 3 most common symptoms with 23 (77%), 19 (63%) and 17 (57%) patients reporting these symptoms, respectively. Chills, ageusia and subjective fevers were the 3 least common symptoms with 8 (27%), 7 (23%) and 5 (17%) patients reporting these symptoms, respectively. The box plot in Figure 2.2 details how long each symptom was reported by patients. Only patients who reported the symptom are included in this analysis. Based on the median number of days, nasal/sinus congestion lingered the longest with a median of 15 days followed by ageusia with a median of 11 days. Although ageusia was only reported by 7 patients, the symptom lingered for a longer time compared to other symptoms. Subjective fevers, chills and chest pain were reported for the shortest period each having a median of 1 day. Patients completed the daily symptom tracker an average of 73% of days enrolled in the study. They wore the Garmin device an average of 90% of days enrolled in the study. This indicates that patient compliance with wearing the device was 17% greater than compliance with answering the daily symptom tracker. This statistic demonstrates the higher efficiency of wearable device data for remote monitoring and helps

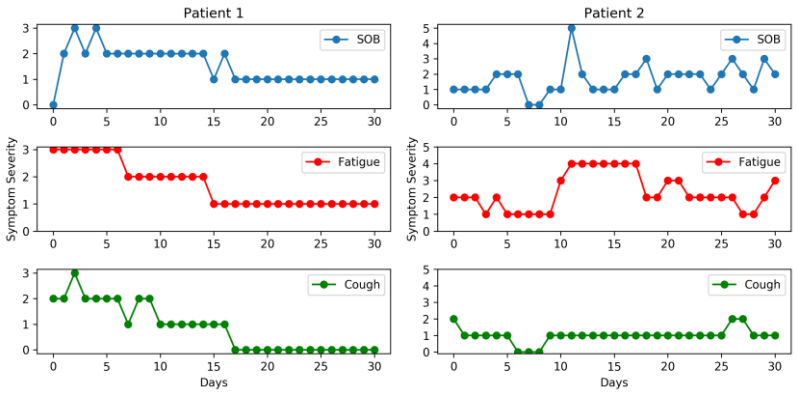

Figure 2.3 Symptom severity progression for two COVID-19 patients. Patient 2's symptom severities decrease by day 7 and then sharply increase again after day 10. The shortness of breath (SOB), fatigue, and cough severities correspond to questions 3-5 of the symptom tracker.

motivate our proposed ML task for patient recovery classification based on automatedly collected device data, as opposed to relying on manually entered symptom data.

2.3.2 Patient Recovery Classification

The objective of this ML task is to classify whether a patient has recovered from COVID-19 symptoms based on their device data. This binary classification model can provide healthcare workers with automated insights into the recovery status of their infected patients and bypass the need for manual daily symptom tracking which relies on patient compliance. To the best of our knowledge, there is no clear definition for full recovery from COVID-19. The US CDC recommends removal of isolation for COVID-19 infection when a patient's symptoms have significantly improved, they have been afebrile for at least 24 hours in the absence of fever-reducing medications, and it has been at least 10 days since symptom onset [62]. However, it is now understood some patients can suffer from ongoing symptoms from COVID-19 for weeks and even months [63]. Unlike symptom severity which can be identified by patients themselves, recovery is a gradual, subtle, and implicit process. In this task, we classify whether a patient has recovered from the COVID-19 symptoms collected by our symptom tracker app. Most patients

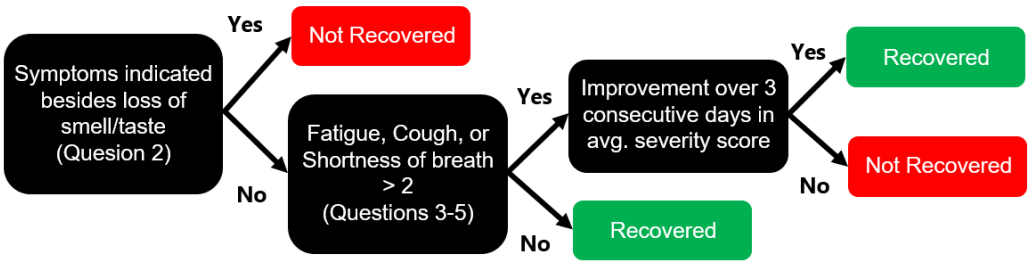

Figure 2.4 Labeling logic for patient recovery classification based on symptom tracker questionnaire responses.

experienced a steady decline in symptom severities, however, some patients initially appeared to recover and then had symptoms re-appear. Figure 2.3 displays the symptom severity progression for the first 30 days for two different COVID-19 patients in terms of shortness of breath (SOB), fatigue, and cough. Patient 1 is an example of a patient who experienced a steady recovery. Patient 2, however, demonstrates a complicated symptom progression. The symptom severities for this patient declined by day 7 and then sharply worsened after day 10, especially for SOB and fatigue. All three symptoms linger for this patient for over a month.

A binary label is generated on a daily basis for each patient: recovered (0) or not recovered (1). The labelling logic for patient recovery was developed in collaboration with UCSD Health doctors and is displayed in Figure 2.4. If symptoms are present besides loss of taste/smell (Question 2), label as not recovered (1). We do not consider loss of smell/taste because these symptoms have been shown to linger after a patient has recovered from COVID-19 [64]. If no symptoms are marked for Question 2 and fatigue/cough/shortness of breath severity is ≤ 2 (Questions 3-5), label as recovered (0). If fatigue/cough/shortness of breath severity is > 2 but there is an improvement over 3 consecutive days in severity scores, label as recovered (0). In order to accommodate for complex cases such as Patient 2 in Figure 2.3, in which there may be a day labeled as recovered (0) between days labeled as not recovered (1), we apply the following logic. If a patient is labeled as recovered (0) for 7 consecutive days, all the following labels are also

Table 2.3 Statistics for label count per patient.

	Mean	Std.	Max	Min	Median
Not Recovered	24	29	85	0	16
Recovered	21	26	76	0	16

marked as recovered (0). Otherwise, the recovered (0) days shorter than 7 days are reverted to non-recovered (1) days. This ensures there are no "recovered" days between "not recovered" days and vice versa. The statistics of the symptom tracker labels are shown in Table 2.3. The average number of "not recovered" and "recovered" samples per patient is 24 and 21, respectively. The median number of "not recovered" and "recovered" samples per patient is 16 for both. This difference in mean and median is the result of outlier patients who have a high amount of one label. There are 10 patients for which 90% of their labels are either "not recovered" or "recovered". Patients with few "not recovered" labels may be a result of being asymptomatic or a delay in joining the study after being infected and testing positive. Patients with few "recovered" labels remained symptomatic for the study duration. These labels are used for the patient recovery classification task. Note that the recovery classification technique proposed here can be used with any other labeling logic developed by other health care providers.

2.3.3 Device Data and Preprocessing

The Garmin vivosmart4 includes a heart rate monitor, accelerometer, ambient light sensor, and blood oxygen saturation (SpO2) monitor. The device uses these sensors in order to calculate various health parameters, including lifestyle and vitals information. The device data is presented in Table IV. The Garmin API documentation provides a description of these parameters [61]. Lifestyle features include activity (steps, distance, floors, active time, etc.), stress (average stress,

Table 2.4 List of Garmin device features that our approach uses. Features marked with * require additional processing after receiving the data from Garmin. Features marked with ^ are available in the dataset from [51] which we discuss in Sec. 2.4.2.

Features
Steps^, Distance, ActiveTime, ModerateIntensityDuration, VigorousIntensityDuration, FloorsClimbed, AverageStressLevel, MaxStressLevel, StressDuration, RestStressDuration, ActivityStressDuration, LowStressDuration, MediumStressDuration, HighStressDuration, SleepDuration^, BedTime*^, UpTime*^, DeepSleepDuration^, LightSleepDuration^, REMSleepDuration^, AwakeDuration^, MinHeartRate^, MaxHeartRate^, MeanHeartRate^, RestingHeartRate^, MinSpO_2*, MaxSpO_2*, MeanSpO_2*

max stress, stress duration, etc.), sleep timing (duration, bed time, up time), and sleep stages (deep, light, REM, awake). Stress-related features are derived based on heart rate variability [61]. The variable length of time in between each heartbeat is regulated by the body's autonomic nervous system. The less variability between beats equals higher stress levels, whereas the increase in variability indicates less stress. As mentioned in the introduction, the researchers in [51] found that COVID-19 affected the number of daily steps and time asleep for patients in their study. This result motivates us to include all lifestyle features when training our patient recovery classification model. In addition to lifestyle factors, the vivosmart4 measures vitals data including heart rate and SpO2. The device is capable of manual SpO2 spot checks during the day and 4 hours of continuous measurement during sleep. Since the symptoms data and patient recovery classification labels are generated daily, we aggregate the device data features for each day. The Garmin Health API provides summarized activity, sleep, stress, and heart rate features daily. The features in Table 2.4 marked with a * require additional processing after receiving the data from Garmin. These include BedTime, UpTime, MaxSpO2, MinSpO2, and MeanSpO2. The BedTime and UpTime features are encoded as the number of seconds before or after midnight (e.g., 11:30 PM bed time is encoded as -1800 seconds, 8:00 AM wake time is encoded as 28800 seconds). Since only the continuous SpO2 data is available through the Garmin API, we transform the SpO2 time series each day into

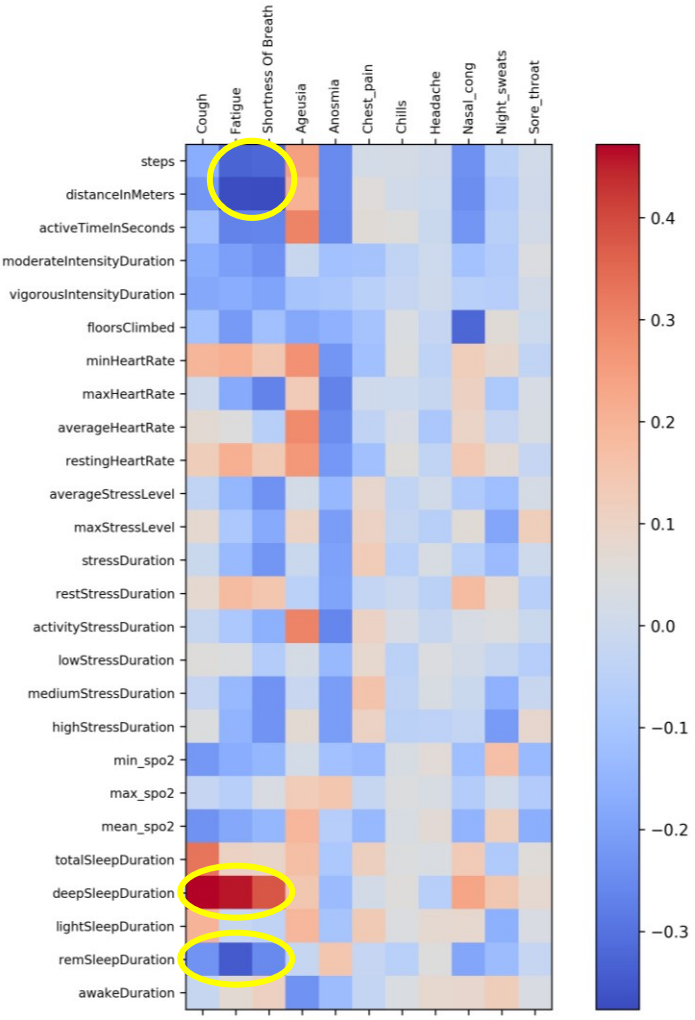

Figure 2.5 Spearman correlation between lifestyle/vitals and symptoms. Notable correlations are circled in yellow.

the MaxSpO2, MinSpO2, and MeanSpO2 features displayed in Table 2.4. Note that a subset of the features is marked with ^ in Table 2.4 indicating they are available in the dataset from [51] which we discuss in Sec. IV (B). Once the device data is aggregated for each day, we match it with the corresponding patient recovery label to form patient-day samples. Each patient-day sample consists of the recovery label and the summarized lifestyle and vitals features for one patient's day in the study. Note that symptoms data are not directly used as part of the training data, but rather to generate the daily patient recovery labels.

Table 2.5 Top 10 correlations between symptoms and device features.

Symptom	Device Feature	Spearman Correlation
Cough	DeepSleepDuration	0.47
Fatigue	DeepSleepDuration	0.46
SOB	DeepSleepDuration	0.38
SOB	DistanceInMeters	-0.38
Fatigue	DistanceInMeters	-0.37
Fatigue	REMSleepDuration	-0.34
Cough	TotalSleepDuration	0.33
Fatigue	Steps	-0.33
SOB	Steps	-0.32
Nasal Congestion	FloorsClimbed	-0.32

Figure 2.5 displays a heatmap of the correlation between the aggregated daily lifestyle/vitals features and symptoms data for our study cohort. We use Spearman correlation because the symptom variables are not continuous. Spearman evaluates the monotonic relationship between two continuous or ordinal variables [65]. The color of each heatmap square describes the magnitude and directionality of the correlation. Darker red squares correspond to a stronger positive correlation while darker blue squares correspond to a stronger negative correlation. Table 2.5 displays the top 10 most significant correlations between symptoms and device features and in Figure 2.5 we circle notable correlations in yellow. These include distance and steps vs. fatigue and shortness of breath (SOB) severity, and deep and REM sleep vs. cough and fatigue severity. The correlations for distance vs. SOB and fatigue are -0.38 and -0.37, respectively. The correlations for steps vs. SOB and fatigue are -0.32 and -0.33, respectively. It is sensible that distance and steps are negatively correlated with cough and SOB severity. A patient is less likely to be active if their symptom severities are higher. Deep and REM sleep duration are positively

and negatively correlated, respectively, with cough, fatigue and SOB severity. The most significant correlation is deep sleep vs. cough, which has a correlation of 0.47. REM sleep is most correlated with fatigue, with a correlation of -0.34. According the American Academy of Sleep Medicine, as the immune system fights infection, the amount of time spent in REM sleep is decreased while deep sleep is increased [66]. This is because it is during deep sleep that many reparative bodily processes occur. This validates the directionality of the correlations between REM/deep sleep and symptom severities. While the individual correlations between other lifestyle/vitals features and symptoms are not as prominent, the heatmap in Figure 2.5 indicates that a combination of these features can provide useful information about symptom severity when training the ML model. Overall, these correlation observations help motivate our ML approach to patient recovery classification based on device data.

2.3.4 Random Forest and Personalization

We train multiple ML classifiers in order to determine which is most effective at modelling the patient recovery task, as described in Sec. IV (A). As indicated in Table 2.6, the Random Forest (RF) model results in the best performance during LOSO CV. In this section, we discuss the operation of the RF model and our personalization technique.

RF is an ensemble model that aggregates a collection of decision trees in order to reduce overfitting and the resulting high variance in prediction [67]. To do this, RF utilizes bootstrap aggregation (bagging) and feature bagging. RF produces bootstrap datasets that are randomly and independently drawn with replacement from the training dataset. Each bootstrap dataset has the same size as the original training set and is used to train a decision tree. Bootstrap aggregation in RF averages the prediction of all decision trees which greatly reduces the variance compared to a single decision tree. Moreover, since individual trees generated in the bagging process are

identically distributed, the expected prediction of RF is the same as the expected prediction of individual trees. Combining the above facts, RF has a lower variance than individual trees, while its bias remains the same [68]. RF further reduces the correlation between its member decision trees by introducing feature bagging, which randomly selects a subset of features when constructing each tree. In addition, RF is known to perform well even when using redundant or irrelevant features. Since we utilize multiple lifestyle and vitals features for model training, it is possible that some features do not provide useful information. Since RF is more robust to noisy features as compared to the other models [69], redundant or irrelevant features will not greatly impact performance.

Multiple studies that focus on ML for health applications have shown that model personalization is a key step in improving performance due to the physiological differences between patients [70-73]. In this study, we observe that vitals and lifestyle factors vary among patients and propose a RF-based personalization technique to tune the model to each patient. Our technique involves including the first *k* days of labeled data from the test patient in the training set. In the traditional RF bootstrapping process, each training sample has uniform weight, which means each data sample is resampled with the same probability. To emphasize the test patient's calibration samples during model training, we assign a greater weight to these *k* samples using the Weighted Bootstrapping algorithm [74]. In order to implement this algorithm, a vector of sample weights $\boldsymbol{W} = \boldsymbol{w_1}, \boldsymbol{w_2}, \ldots, \boldsymbol{w_N}$ is maintained where *N* is the total number of training samples. Weights $\boldsymbol{w_1}, \ldots, \boldsymbol{w_k}$ correspond to the *k* personalization samples from the test patient and are given larger values. Weights $\boldsymbol{w_{k+1}}, \ldots, \boldsymbol{w_N}$ correspond to the data samples from the remaining patients used for training and are assigned lower values. The operation of the Weighted Bootstrapping algorithm is as follows [74]: First, a new bootstrap dataset for one decision tree is initialized. Then,

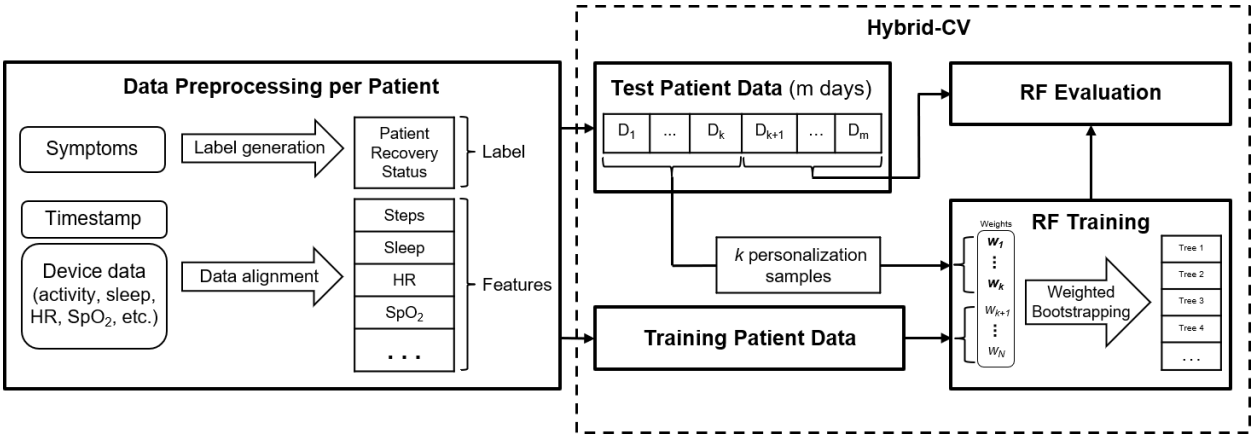

Figure 2.6 Block diagram of our proposed RF personalization approach. After data preprocessing, the first k samples from the test patient are included in the training set during Hybrid-CV. These samples are assigned larger weights, which are bolded in the figure, during weighted bootstrap aggregation.

the weights in $\boldsymbol{W}$ are mapped into the interval $\left[\mathbf{0},\ \sum_{j=1}^{N} \boldsymbol{w_j},\right]$ with subintervals $\boldsymbol{I_1, I_2, \ldots\ I_N}$. The length of each subinterval is proportional to the value of its weight. Next, each data sample is drawn using subintervals $\boldsymbol{I_1, I_2, \ldots\ I_N}$ and the uniform distribution function. The process repeats *N* times such that the size of all bootstrap datasets equals that of the original dataset. Consequently, the samples with higher weights are more likely to appear in each bootstrap dataset. In Sec. IV (B), we compare the performance for different values of *k* and different values of $\boldsymbol{w_1, \ldots, w_k}$. Figure 2.6 displays a block diagram of our proposed RF personalization technique. After preprocessing each patient's data, Hybrid-CV is carried out in which the training and test sets are split on a per patient basis and the first *k* days of test patient data are added to the training set as personalization samples, as shown in Figure 2.6. These *k* samples are assigned greater weights, which are bolded in the figure, during weighted bootstrapping. After training, the model is evaluated on the remaining, future data samples of the test patient.

2.4 Results and Discussion

In this section, we describe the experiment settings and present patient recovery classification results. We discuss the effects of our RF model personalization technique on performance and carry out feature analysis using Shapley Values in order to interpret what the model has learned. Finally, we provide a discussion on the challenges encountered during this study.

2.4.1 Experiment Setting

We implement and evaluate our machine learning models using the Scikit-learn library in the python environment on an Intel i5 3.2GHz quad-core and 16GB RAM computer. Accuracy, sensitivity, specificity, and F1-score are calculated and used as our evaluation metrics for the patient recovery classification task. For this task, a negative and positive sample correspond to a "recovered" and "not recovered" patient-day sample, respectively. Accuracy returns an overall measure of how much the model is correctly predicting on the entire set of test data. Sensitivity and specificity measure the true positive and true negative rate, respectively. F1 score is calculated as the harmonic mean of precision and recall (sensitivity) and is used to find the best trade-off between the two quantities [75]. As a result, we use F1 score for deciding the top performing model.

We carry out LOSO CV to mirror the clinically relevant use-case scenario of diagnosis for newly infected subjects [82]. LOSO CV separates the data into train and test sets on a per patient basis in order to simulate the practical application. This data split ensures that data from the same patient does not appear in both the training and testing sets. We use LOSO CV to compare the performance of different ML models. We then carry out Hybrid-CV, in which a specified number of samples from the test patient are included in the training set. These personalization samples are

not included in the test set to ensure there is no overlap between train and test sets at the sample level. We compare how performance is affected by applying varying levels of personalization using our RF-based personalization technique described in Sec. III (D). Since the number of samples for each patient is different based on their participation in the study, the training and testing sets will vary in size for both CV experiments. Instead of averaging the results over each data split, we save the model predictions for each data split and calculate metrics over all predictions. This ensures that each patient-day contributes equal weight to the final result.

In the LOSO CV experiment, we compare RF with the following ML models: logistic regression (LR) [76], k-nearest neighbors (KNN) [77], support vector machine (SVM) [78], artificial neural network (ANN) [79], and long short-term memory (LSTM) neural network [80]. Model hyperparameter tuning is performed with each training set using a randomized search over a predefined hyperparameter grid for each model. Since LSTM models take sequential data as input, we organize the lifestyle and vitals features into sequential data samples using a window length of 7 days and a step size of 1 day. A step size of 1 day is used to extract the maximum number of samples. As a result, each input sample has a dimension of (7, Nfeatures) where Nfeatures represents the number of lifestyle and vitals features. The patient recovery label for the last day of each window is assigned to each input sample. We train the LSTM as a many-to-one model, as opposed to a many-to-many model, since the application of this method is only concerned with estimating whether the patient is recovered or not for the current day. In addition, training the LSTM to estimate one label at a time matches the process for the other ML models, resulting in a fairer comparison. We carry out two LSTM experiments using 16 and 32 hidden units for the LSTM layer followed by a fully connected layer with 1 output unit. For these experiments, we train the models using the Adam optimizer [81] and a dropout rate of 50% to

Table 2.6 Comparison of ML model performance for LOSO CV.

Model	Acc	Sens	Spec	F1
LR	0.60	0.61	0.52	0.61
ANN	0.59	0.62	0.62	0.63
SVM	0.54	0.61	0.59	0.62
KNN	0.55	0.51	0.68	0.60
LSTM-16	0.63	0.56	0.71	0.61
LSTM-32	**0.64**	**0.64**	0.60	0.64
RF	0.59	0.52	**0.78**	**0.66**

reduce overfitting. For the LSTM layers, we use a sigmoid activation function for the input, forget and output gates, and a hyperbolic tangent (tanh) activation function for the cell state and hidden state. The fully connected layers use a sigmoid activation function and we use binary cross entropy loss as the loss function. We experimented with different numbers of training epochs and batch sizes and found that 25 epochs and a batch size of 32 resulted in the best performance.

2.4.2 Patient Recovery Classification Results

Accuracy, sensitivity, specificity, and F1-score for each ML model during LOSO CV are presented in Table 2.6. The LSTM-32 model achieves the highest accuracy and sensitivity, both equal to 0.64, while the RF model achieves the highest specificity and F1 score equal to 0.78 and 0.66, respectively. As described in the experiment setting, we use F1 score for deciding the top performing model since this metric calculates the tradeoff between precision and sensitivity. Since RF achieves the highest F1 score, we conclude that RF is the best performing model for patient recovery classification. We attribute the RF's top performance to its ability to reduce the variance in prediction via the bagging process and its robustness to redundant or irrelevant features. The LSTM-32 model is the second-best performing model, indicating that meaningful temporal information exists in the data for estimating recovery from COVID-19. Since RF is the top

Table 2.7 Hybrid-CV results using different levels of personalization.

Personalization Samples	Acc	Sens	Spec	F1
0	0.59	0.52	0.78	0.66
1	0.63	0.59	0.75	0.70
2	0.67	0.66	0.71	0.75
3	0.72	0.73	0.68	0.79
4	0.80	0.86	0.64	0.86
5	0.82	0.89	0.63	0.88

performer, we use this model in the next experiment to understand how the number of personalization samples impacts RF performance.

Next, we discuss the results of the Hybrid-CV experiment. As mentioned in the experiment settings, LOSO CV separates the data into train and test sets on a per patient basis. Since physiology and lifestyle differ between patients, we apply varying levels of personalization during the Hybrid-CV experiment. We implement our RF-based personalization technique by including the first 1-5 days of test patient data in the training set. These personalization samples are assigned a larger weight so that they are sampled more frequently during the bootstrap aggregation step. Table 2.7 displays the results for different amounts of personalization. Evidently, the classification results are worse when no personalization is applied. The accuracy, sensitivity, specificity, and F1-score are 0.59, 0.52, 0.78, and 0.66, respectively, when no personalization is applied. As personalization samples are included in the training set, accuracy, sensitivity, and F1-score increase, while specificity decreases. When using 5 personalization samples, the accuracy, sensitivity, specificity, and F1-score are 0.82, 0.89, 0.63, and 0.88, respectively. Since the personalization samples for each patient correspond to their first 1-5 days in the study, these samples are primarily labeled 1 or “not recovered”. This means that as more personalization samples are included in the training set, the model can increasingly learn the infected baseline of

Table 2.8 Performance comparison when applying different RF bootstrap aggregation weights to 5 personalization samples.

Bootstrap Aggregation Weights	Acc	Sens	Spec	F1
1	0.70	0.69	0.73	0.77
10	0.82	0.89	0.63	0.88
100	0.81	0.88	0.62	0.87

the patient based on their vitals and lifestyle data. This causes the sensitivity to increase since the model will be able to increasingly correctly classify a patient who has not recovered. This corresponds to increasing true positives (classifying a patient as not recovered when they are indeed not recovered) while minimizing false negatives (classifying a patient as recovered when they are not recovered). As the sensitivity increases, the specificity decreases. Since the model is increasingly tuned to classify a patient as not recovered, this will result in more false positives and a lower specificity. For this ML task, false positives are more acceptable than false negatives. Classifying a patient as not recovered when they are recovered is less harmful than classifying a patient as recovered when they are not recovered. Overall, adding personalization samples increases the model performance. When applying this personalization technique to a new patient, the first few days will involve data collection without any classifications from the ML model. After this initial data collection, the personalized model will provide estimations with improved accuracy, sensitivity, and F1-score. The results demonstrate the potential for ML-assisted remote patient monitoring to supplement traditional manual monitoring tools, like daily manual symptom tracking.

The results presented in Table 2.7 are generated by setting the bootstrap aggregation weights for the personalization samples to 10. This means these samples are 10 times more likely to be sampled during the RF weighted bagging process. In Table 2.8, we compare how

Table 2.9 Evaluation of proposed method on open dataset from [51].

	Acc	Sens	Spec	F1
W/O Personalization	0.49	0.33	0.73	0.44
W/ Personalization (5 samples)	**0.61**	**0.55**	**0.67**	**0.61**

classification performance is affected by applying different bagging weights to 5 personalization samples. We set the weights to 1, 10 and 100. Using a bagging weight of 1 means the personalization samples have the same probability of being sampled as the training data from other patients. Evidently, a bagging weight of 1 produces worse performance with an accuracy, sensitivity, specificity, and F1-score of 0.7, 0.69, 0.73, and 0.77, respectively. In this case, the personalization samples are not emphasized and the model is not effectively calibrated. Increasing the bagging weight from 10 to 100 does not improve model performance. This indicates that at a certain weight, the personalization samples are sampled frequently enough during bagging to effectively calibrate the model. Further increasing the bagging weight does not provide additional utility in model personalization.

In order to extend the evaluation of our proposed method, we applied our approach to the dataset collected in [51]. This dataset includes sleep, heart rate and steps data collected from a wearable device, and the date of first symptoms and date of recovery which are manually recorded by each patient. Since this dataset does not include SpO2, stress or activity (besides steps) data, the number of features is significantly less than our own dataset (12 vs. 28). In Table 2.4, features marked with ^ are available in the dataset in [51]. We labelled all days between the start of symptoms and recovery dates as “not recovered” and all days after the recovery date as “recovered.” We then combined these labels with the corresponding device features to create the dataset in the same manner as our experiment setting. After these data processing steps, 15 patients

had sufficient data to be included in this experiment. Table 2.9 displays the results when applying our method to this dataset. We train a Random Forest model with and without personalization and calculate the accuracy, sensitivity, specificity, and F1-score. We use 5 samples when applying our personalization technique and observe that the performance significantly improves compared to the non-personalized results. With personalization, our approach achieves an accuracy, sensitivity, specificity, and F1-score of 0.61, 0.55, 0.67, and 0.61, respectively. Evidently, the performance metrics are not as good for this dataset. This may be due to the limited feature set and inaccurate recovery dates recorded by patients. We observe similar patterns in the results compared with our own dataset which include that there is a performance enhancement when applying our personalization technique. Overall, these consistent observations between our dataset and the dataset in [51] indicate that our proposed approach is not only applicable to our dataset, but can potentially be applied to different datasets collected in clinical practice.

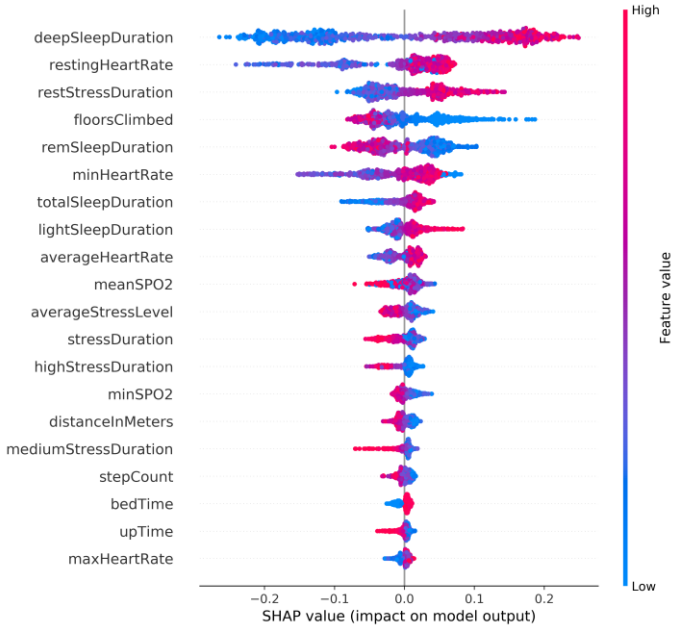

Figure 2.7 Summary of Shapley top features where each point corresponds to a data sample. The x-axis represents a feature's impact on model output. Positive SHAP values push the model to output 1 or "not recovered".

2.4.3 Model Interpretability via Shapley Value Analysis

Next, we utilize Shapley Values [83, 84] in order to determine which lifestyle and vitals features have the most significant effect on model classification for our dataset. Shapley Value analysis is a model-agnostic interpretation method derived from game theory. Given a set of feature values and a trained machine learning model, the estimated Shapley value indicates how each feature contributes to the model's classification. We use the tree SHAP (SHapley Additive exPlanations) framework [85, 86], which is optimized for tree-based models, to interpret the output of the RF model for patient recovery classification. Figure 2.7 displays the Shapley results where the features are ranked from the top to bottom based on their impact on the model's output. Each point on the plot corresponds to an individual data sample and represents the contribution from the feature listed on the Y-axis to the RF's classification. The placement on the X-axis represents the amount of positive/negative contribution to the classification. Positive contribution corresponds to pushing the model to estimate that a patient is not recovered. The color of each point represents the actual value of the feature (red is high while blue is low). The top two features based on Shapley analysis include deep sleep duration and resting heart rate. Higher values of deep sleep duration (colored in red) contribute to a positive, or not recovered, classification. This observation aligns with the correlation analysis presented in Sec. 2.3.3. As mentioned earlier, deep sleep increases when a patient is sick since this is when many reparative bodily processes occur. Increased resting heart rate also contributes to a positive classification by the RF model. This relationship makes sense since resting heart rate will decrease as a patient recovers. Additional observations include that a lower number of floors climbed contributes to a positive classification while an increased mean SpO2 contributes to a negative, or recovered, classification. Both relationships are sensible,

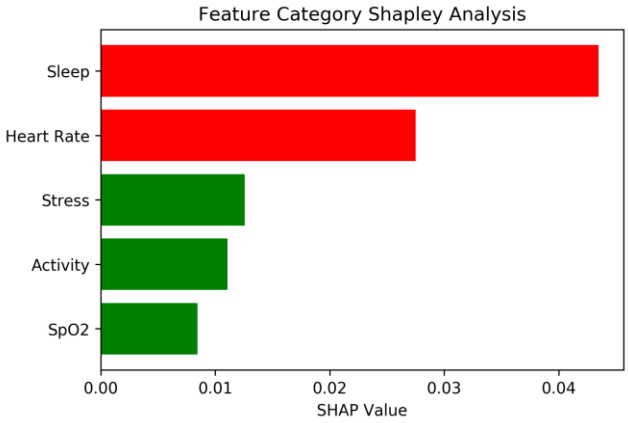

Figure 2.8 Impact of feature categories on model output. Features are grouped into 5 categories and a categorical SHAP score is calculated. Red or green bars indicate that an increase in the category's feature values pushed the model to output "not recovered" or "recovered," respectively.

as a patient who has not recovered will be less active and a patient who has recovered will have a higher SpO2.

In addition to analyzing the impact of individual features, we grouped the features into 5 categories (Activity, Sleep, Stress, Heart Rate and SpO2) and investigated their impact on model output. A SHAP score for each category was calculated as the average of the absolute SHAP values for the features in that category. Figure 2.8 displays the ranking of feature categories based on their categorical SHAP score. We also examined whether, on average, an increase in the feature values for each category pushed the model to estimate "recovered" or "not recovered." In Figure 2.8, a red colored bar indicates that an increase in the category's feature values pushed the model to output "not recovered." A green colored bar indicates that an increase in the category's feature values pushed the model to output "recovered." Evidently, the sleep category had the most significant impact on model output. An increase in feature values in the sleep and heart rate categories pushed the model to estimate "not recovered" (red bars) while an increase in feature values in the stress, activity and SpO2 categories pushed the model to estimate "recovered" (green bars). Overall, the individual feature and feature category Shapley analysis demonstrates that our

model can learn clinically relevant relationships between device data and the status of patients. The interpretability of a ML model is necessary for humans to understand what the model has learned, especially in medical applications.

2.4.4 Limitations and Research Challenges Encountered

In this section, we discuss limitations to our proposed approach and challenges faced while implementing this study. One limitation in our approach is that patients were only enrolled and provided devices for data collection after testing positive for COVID-19. It is likely that some patients started experiencing symptoms before going for a COVID-19 test. This meant we were not able to collect symptoms and wearable data during the initial days of the infection. In order to ensure that data can be collected before and during the onset of COVID-19 infection, participation could be made available to a larger number of patients that already own a wearable device. After testing positive for COVID-19, a patient could immediately enroll and begin sharing both past and current data. Another limitation to our approach is that the RF model does not process data sequentially while the progress of COVID-19 is sequential. In this work, we experimented with LSTM, a popular temporal model, however, found its performance to be worse than RF. Training an LSTM requires significantly more data since neural networks are highly prone to overfitting when the underlying dataset size is small [87, 88]. In order to fully utilize temporal relationships in the data, we plan to further investigate sequence modeling with additional data in our future work. This will include implementing many-to-many sequence models using different time windows to learn temporal progression along with the label. In addition, a larger dataset can enable the use of additional features such as patient demographic information. The model may learn relationships between COVID-19 recovery and demographic data such as age, gender, and ethnicity.

Concern over privacy was an issue encountered during recruitment for this study. As mentioned in Sec. III (A), we recruited patients from both the UCSD Health and Neighborhood Healthcare (NH) COVID-19 telemedicine clinics. NH is a community clinic that primarily provides care to underserved populations. In order to increase accessibility to our study, we developed a Spanish version of our symptom tracker app with assistance from NH. Overall, we experienced more difficulty recruiting from NH. One reoccurring reason why NH patients did not want to partake in our study included a concern over privacy. Certain patients expressed discomfort over wearing the device 24/7 due to concerns of being tracked. Our recruitment personnel would highlight that the device does not collect any location data, however, certain patients still declined participation. The above challenge encountered during our study showed that privacy concerns and lack of trust in wearables may further limit access and use of digital technologies by underserved populations, contributing to an increased digital divide in healthcare. As healthcare begins to rely more on digital technologies, these concerns must be addressed in order to ensure equal access to high quality healthcare [89].

2.5 Conclusion

In this chapter, we propose an intelligent remote monitoring platform, namely eCOVID, for enhanced COVID-19 ambulatory care. Based on data collected from our study with the UCSD Health and Neighborhood Healthcare COVID-19 telemedicine clinics in San Diego County, we demonstrate correlations between automatically collected wearable data and manually entered symptom data. We propose a novel ML approach to estimate whether a patient has recovered from COVID-19 symptoms based on the automatically collected wearable data. Our results demonstrate that ML-assisted remote monitoring using wearable data can supplement or be used in place of manual daily symptom tracking which relies on patient compliance.

By developing and demonstrating the ability to track patient recovery status remotely, our approach can enable more optimal care of COVID-19 ambulatory patients at scale. Care teams will be able to track patient recovery efficiently through automatically generated and updated dashboards instead of the current practice of manual symptom tracking and phone calls, the latter becoming ineffective when there is a surge in cases. This shift can lead to significant improvement in the efficiency and scalability of ambulatory patient care, while at the same time enabling savings in human and equipment resources. Moreover, the approach can be used for providing scalable and efficient care for future pandemic and epidemic challenges.

In the next chapter, we present the results of a single-arm nonrandomized trial which assessed the effectiveness of a fully digital, autonomous, and ML–based lifestyle coaching program on achieving BP control among adults with hypertension. The study demonstrates that the ML–based lifestyle intervention helped hypertension patients achieve meaningful BP improvements and high engagement, while substantially reducing clinician workloads.

Chapter 2, in part, is from the material as it appears in the IEEE Journal of Biomedical and Health Informatics, 2023, Leitner, Jared; Alexander, Behnke; Chiang, Po-Han; Ritter, Michele; Millen, Marlene; Dey, Sujit. The dissertation author was the primary investigator and author of this paper.

Chapter 3 The Effect of an AI-Based, Autonomous, Digital Health Intervention Using Precise Lifestyle Guidance on Blood Pressure in Adults with Hypertension: Single-Arm Nonrandomized Trial

3.1 Introduction

3.1.1 Background

High blood pressure (BP), or hypertension, is one of the most prevalent chronic diseases in the world [90]. Hypertension affects 48% (approximately 120 million) of adults in the United States, and 78% (approximately 93 million) of the cases are uncontrolled (ie, BP≥130/80 mm Hg) [92]. Hypertension is a major risk factor for stroke and acute myocardial infarction [93] and remains a large public health challenge with an extra cost of US $2000 per year per hypertension patient, resulting in an additional US $131 billion in annual health care costs in the United States [94]. The American College of Cardiology and American Heart Association's clinical practice guidelines define hypertension as systolic BP (SBP)≥130 mm Hg or diastolic BP (DBP)≥80 mm Hg, consistently over time [91]. A large-scale analysis of 48 randomized clinical trials showed that a 5 mm Hg reduction in SBP lowered the risk of major cardiovascular events by 10% [95], highlighting the importance of developing new strategies to achieve hypertension control at scale.

Hypertension management typically begins with home monitoring of BP to gain a more accurate estimate of a patient's BP within their usual, daily routine [96]. However, self-monitoring without additional support is not associated with lower BP or better control [97-99]. Lifestyle management in conjunction with self-monitoring is effective in controlling BP as lifestyle factors (eg, activity, sleep, diet, and stress) have a substantial impact on BP [100-103]. Even for patients taking antihypertensive medication, lifestyle management can enhance medication efficacy, leading to better BP control [104]. Traditionally, lifestyle management involves patients with hypertension visiting their primary care physician (PCP) and receiving guidance on lifestyle

modifications that are generally known to improve BP. However, due to time constraints related to workload, physicians are often unable to optimally counsel patients on lifestyle modifications or personalize their guidance [105,106]. Due to insufficient guidance and the lack of feedback in between clinic visits, patients may implement some of these changes; however, patient engagement and compliance are generally suboptimal for achieving control. To improve patient engagement, new digital health technologies and remote patient monitoring programs have been developed for hypertension care [107-110]. These programs typically provide patients with remote monitoring devices (eg, BP cuffs and activity trackers) and match patients with health coaches. BP and lifestyle data collected from remote monitoring devices allow health coaches to view trends and make personalized recommendations to patients. However, these approaches do not consider the individual impact of lifestyle factors on BP, which may vary across individuals due to physiological differences. Furthermore, the reliance on health coaches is highly time and resource intensive, resulting in a high operating cost, which significantly limits scalability [111].

3.1.2 Objectives

To address the challenges of poor patient engagement due to generic, insufficient guidance and limited scalability of care due to human coaching models, we propose an artificial intelligence (AI)–driven, autonomous, precise lifestyle coaching program for patients with hypertension. The intervention platform consists of a monitoring system that ingests lifestyle and BP data and builds personalized machine learning (ML) models to determine the individual impact of different lifestyle factors on BP. On the basis of the lifestyle impact analysis, the system autonomously provides precise lifestyle recommendations delivered to a patient's smartphone that enable patients to focus on specific aspects of their lifestyle that have the greatest associations with their BP. While the platform autonomously engages patients, it is clinician supervised and notifies clinicians of

critical BP readings. In our previous study [112], we enrolled 38 participants who were prehypertensive or had stage 1 hypertension (SBP between 120 and 139 mm Hg or DBP between 80 and 89 mm Hg) and demonstrated that 75% of the participants receiving the intervention were able to achieve a controlled BP (<130/80 mm Hg) after 16 weeks of engagement. However, the limitations of the previous study [112] are as follows: (1) the participants were not provided with an interactive mobile app for the delivery of our precise lifestyle recommendations, (2) the small number of participants did not enable rigorous evaluation, and (3) the study did not consider patients with stage 2 hypertension who can potentially benefit more from lifestyle management.

This study aims to evaluate the effectiveness of our AI-based, precise lifestyle guidance coaching program in helping patients with stage 2 hypertension achieve BP control and demonstrate the platform's scalability. The primary study objectives are to evaluate the change in BP and the percentage change of participants in different BP categories (controlled, stage 1 hypertension, and stage 2 hypertension) over time (baseline, 12 weeks, and 24 weeks). Secondary objectives include assessing participant engagement as measured by consistency of data collection and interactions with our mobile app and determining the number of manual clinician interventions, as defined by the escalation rules set for the study, to assess the potential scalability of our approach.

3.2 Methods

3.2.1 Recruitment

This study was performed in collaboration with the University of California, San Diego Health's Population Health Services Organization (PHSO). Participants were enrolled on a rolling basis from November 2021 to February 2023. The inclusion criteria required participants to have stage 2 hypertension (SBP≥140 mm Hg or DBP≥90 mm Hg per the American College of

Cardiology and American Heart Association's 2017 guidelines [91]) based on their most recent clinical measurements and to be fully ambulatory (ie, not requiring an assistive device such as a cane, wheelchair, or walker). In addition, participants were required to be aged ≥18 years at enrollment, English speaking, and own an Android or iPhone (Apple Inc) smartphone. The trial was designed in a fully remote manner so that participants could participate entirely from home. The PHSO care team aggregated a list of patients who met the inclusion criteria and sent a recruitment flyer via bulk message using the Epic MyChart (Epic Systems Corporation) messenger. The flyer introduced the study and instructed patients to email the study team if they were interested in participating. After contacting the study team, eligible patients were asked to complete an electronic informed consent form. Patients who consented were sent a Fitbit Inspire 2 (Fitbit Inc) and a Bluetooth-enabled Omron Silver (Omron Corporation) BP monitor to collect their lifestyle and BP data for up to 6 months. Each shipment included instructions for self-onboarding, which described the steps to set up and connect the devices to the patient's mobile phone. Patients who already owned a Fitbit or Apple Watch (Apple Inc) had the option to use their device instead of receiving one from the study team. Patients who required an extra-large cuff were provided an iHealth Ease (iHealth Labs Inc) BP monitor instead of an Omron Silver.

3.2.2 Ethical Considerations

This study (protocol #181405) was reviewed and approved by UC San Diego's Human Research Protections Program, which operates Institutional Review Boards. All participants in this study provided informed consent, which included the collection of their data and the provision of study results derived from their individual data. The confidentiality and privacy of participants were ensured by assigning a deidentified code to each patient. While participants were not offered

monetary compensation, those without a BP monitor or wearable device were provided with these devices. The study was registered at ClinicalTrials.gov (NCT06337734).

3.2.3 Study Design and Data Collection

We collected data from each participant using a Fitbit or Apple Watch, Omron or iHealth wireless BP monitor, and the study's questionnaire mobile app. Participants were asked to wear their Fitbit or Apple Watch as often as possible, including during sleep, and take 1 to 2 BP measurements per day, in the morning (8 AM-10 AM) or evening (7 PM-9 PM). We provided participants with instructions on how to take accurate resting BP readings [113] and asked that they take 3 consecutive readings during each morning and evening session. This resulted in 1 to 2 sets of 3 measurements per day, and the average of the 3 measurements was used as the final value for each session. Participants synced their BP data to the Omron or iHealth mobile app and their Fitbit data to the Fitbit mobile app; subsequently, the data were automatically uploaded to the Omron, iHealth, or Fitbit clouds. These data were retrieved remotely through the application programming interfaces (APIs) provided by Omron, iHealth, and Fitbit. Data from the Apple Watch were synced with the study mobile appl and uploaded via a custom API to our server. In addition, participants completed a daily questionnaire using our study mobile app that asked about their stress, mood, and dietary choices over the past 24 hours. These questions were developed in collaboration with physicians on our team. The diet questions are tailored to measure information relevant to hypertension, including alcohol, red meat, fruits or vegetables, and salt consumption [118]. The details of the questionnaire are described in our previous study [112]. In addition, we asked participants to complete a study experience survey that asked them to rate the difficulty level of completing the study tasks, how useful they found the recommendations, and their experience

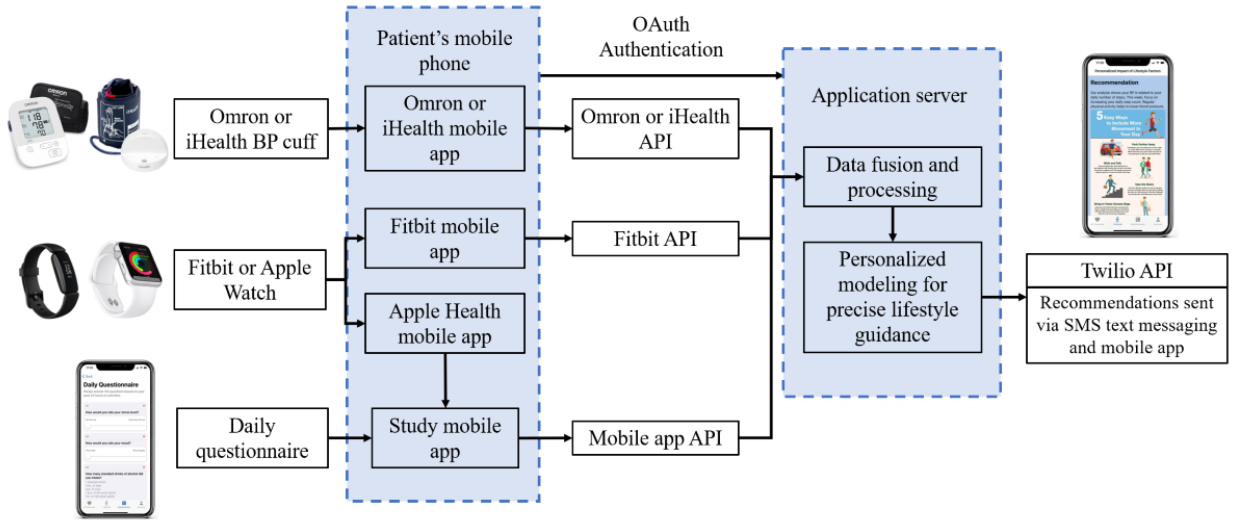

Figure 3.1 Architecture of data transmission. Participant data were collected from Bluetooth-enabled blood pressure (BP) monitors, wearable devices, and a mobile app–based questionnaire. Data were uploaded through the respective application programming interfaces (APIs) to our app server, where the individualized analysis was carried out before delivering recommendations.

using the app. These responses were collected through the mobile app and used to assess participant experience. Figure 3.1 describes the system architecture and data transmission.

Wrist-worn activity and sleep trackers have been widely used in health-related research studies [119], and devices such as Fitbits and Apple Watches have been shown to accurately measure parameters such as step count, heart rate, and sleep duration [120,121]. Fitbits and Apple Watches include an optical heart rate monitor and a 3-axis accelerometer. The devices use these sensors to calculate various health parameters, including lifestyle and vitals measurements. Lifestyle factors include activity (eg, steps, walking and running speed, and active time), sleep timing (eg, sleep duration, bedtime, and uptime), and sleep stages (ie, deep, light, rapid eye movement, and awake). These lifestyle factors are used as part of the intervention, in which we use ML techniques to determine which of the factors have the greatest association with a participant's BP and base our guidance on this analysis.

3.2.4 Description of the Intervention

The intervention is intended to support participants' daily efforts to improve BP and overall cardiometabolic function by facilitating behavioral changes that target physical activity, sleep hygiene, stress management, and dietary choices most relevant to their BP. The intervention platform uses remotely collected lifestyle and BP data to provide personalized, precise, and proactive lifestyle coaching using AI to participants with hypertension. The system integrates the data described in the previous section into a combined data set for each participant. Each participant's personal data set consists of lifestyle features (eg, step count, sleep duration, and salt consumption) that are time aligned with their BP measurements, which serve as the labels for training the ML model. Therefore, each participant's data set is used to train a personal ML model that can predict BP using the participant's lifestyle data as input. With this trained model, the intervention system can determine how different aspects of lifestyle affect the participant's BP. On the basis of the model's determination of the lifestyle factors' impact, the system generates precise lifestyle recommendations. Each lifestyle factor is mapped to a corresponding lifestyle recommendation that was designed with physicians on our team to be consistent with evidence-based clinical guidelines. Furthermore, prior studies have demonstrated that these recommendations, such as increasing step count [123,124], improving sleep quality [125,126], managing stress [127], and reducing salt consumption [128,129], can result in BP reduction. The objective of these precise lifestyle recommendations is to encourage participants to concentrate on 1 aspect of their lifestyle at a time, focusing on the factor that has the greatest association with their BP based on the underlying relationship between their BP and lifestyle factors. We describe the AI-based intervention platform in more detail in our previous study [112].

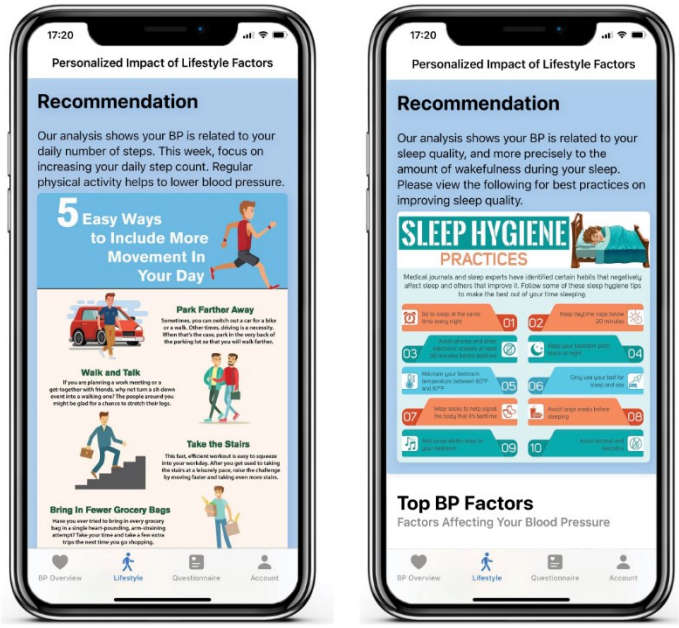

Figure 3.2 Lifestyle recommendations delivered in the mobile app. Participants received weekly lifestyle recommendations based on their data and personalized analytics. The recommendations encouraged participants to prioritize a single lifestyle modification at a time, focusing on the factor that had the greatest impact on their blood pressure (BP).

Participants received weekly lifestyle recommendations based on their data and personalized analytics, which continuously evolved over time. These recommendations were delivered to participants via programmable text messages using the Twilio API (Twilio Inc) service [114] and were displayed in the study mobile app. Each text message included a summary of the participant's BP progression for the current week in addition to the lifestyle recommendation. Figure 3.2 displays examples of these weekly lifestyle recommendations provided in the study app. In addition, patients completed a midweek check-in on the app, which asked whether they could follow each recommendation (yes or no) and to rate the recommendation difficulty on a scale from 1 to 5.

The system includes a safety mechanism to involve clinician intervention in the case of critically high or low BP readings. Critically high BP was defined as SBP>180 mm Hg or DBP>110 mm Hg, and critically low BP was defined as SBP<90 mm Hg or DBP<60 mm Hg [91].

After a critical reading, participants received a text message asking them to remeasure their BP and prompting them to seek assistance or call their medical provider if they were experiencing certain symptoms (eg, chest pain and severe headache). After 2 critical readings in a row, an escalation notification was sent to the PHSO care team via email for manual outreach. To avoid notification fatigue, we limited the number of critically high or low BP notifications sent to the care team to 1 notification per week for a patient.

3.2.5 Primary Outcomes: BP Change and Population Hypertension Control

The first primary outcome was the change in SBP and DBP from baseline to 12 weeks and 24 weeks. A participant's baseline BP was calculated as the average of their readings during the first week of the study. The 12th- and 24th-week BPs were a participant's average reading during that week of the study plus 1 week and minus 1 week. We included BP measurements from 1 week before and after to get a more representative result. For example, the 12-week value was the average of all readings from weeks 11 to 13. As previously mentioned, a 5 mm Hg reduction in SBP can lower the risk of major cardiovascular events by 10% [95]. This motivated us to determine the percentage of participants who experienced >5 mm Hg reduction in SBP at 12 weeks and 24 weeks. To understand the effect on participants with different baseline BPs, we carried out subgroup analysis in which participants were sorted into 3 groups based on their baseline BP: (1) controlled (SBP<130 mm Hg and DBP<80 mm Hg), (2) stage 1 hypertension (SBP 130-139 mm Hg or DBP 80-89 mm Hg), and (3) stage 2 hypertension (SBP≥140 mm Hg or DBP≥90 mm Hg).

Another primary outcome was the percentage change of participants in different BP categories from baseline to 12 weeks and 24 weeks. To assess this, we calculated the percentage of participants who were in the controlled, stage 1 hypertension, and stage 2 hypertension

categories at baseline, 12 weeks, and 24 weeks. Using these percentages, we determined the percentage change from baseline to 12 weeks and 24 weeks.

3.2.6 Secondary Outcomes: Participant Engagement and Clinician Intervention

A secondary outcome measured participant engagement as determined by the consistency of data collection and interactions with our mobile app. The 3 main tasks participants were asked to complete included measuring BP, syncing their wearable device, and answering the mobile app questionnaire. As a result, we used these 3 tasks as our measure of engagement and calculated the percentage of participants completing each of these tasks each week. A participant was marked as engaged for a given week if they provided a BP reading, synced their wearable device data, and answered the questionnaire at least once during the week.

Another secondary outcome was the number of times participants were escalated to the PHSO care team for manual follow-up. The objective of this outcome was to determine the care team's time and resource requirements to implement the intervention and assess the scalability of our approach. The condition for care team intervention was 2 critical BP readings in a row, as previously described.

3.2.7 Statistical Analysis

Descriptive statistics (eg, mean, SD, and percentage) were calculated to describe the demographic and baseline clinical characteristics of the enrolled study population. We compared the characteristics between subgroups based on their baseline BP classification. Change in SBP and DBP from baseline to 12 weeks and 24 weeks was analyzed using a 2-tailed paired Student t test with the level of statistical significance set to $P<.05$. Furthermore, 95% CIs were calculated for these changes. Baseline and follow-up BP data were normally distributed. The McNemar nonparametric test was used to examine the change in the proportion of participants in the

controlled, stage 1, and stage 2 BP range from baseline to 12 weeks and 24 weeks. The McNemar test is used to determine if there is a statistically significant difference in proportions between paired data. We conducted all statistical analyses with Python 3.9 (Python Software Foundation) using the NumPy, Pandas, and SciPy libraries.

3.3 Results

3.3.1 Feasibility Outcomes: Recruitment, Adherence, and Participant Experience

Participants were enrolled on a rolling basis from November 2021 to February 2023. Figure 3.3 details the recruitment numbers and participant flow through the study. A total of 274 patients responded to the Epic MyChart recruitment message by contacting our team and expressing interest. In total, 164 patients consented to join the study out of which 141 (86%) were onboarded and started collecting data. There was a 9.2% drop rate from the start of the study to 12 weeks and

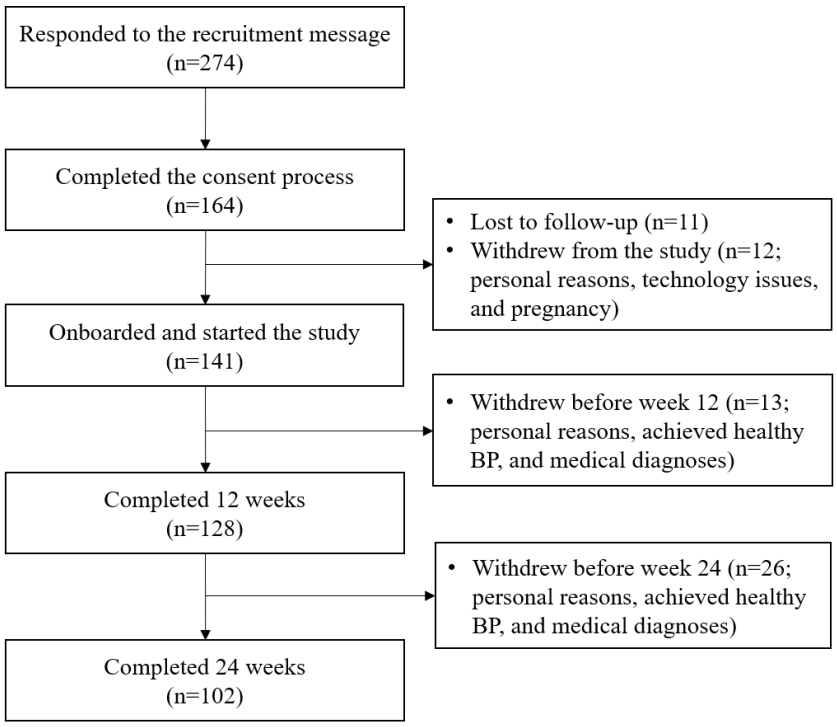

Figure 3.3 Flow of participants through the study. Adults with hypertension were enrolled from the University of California, San Diego Health between November 2021 and February 2023 into a single-arm nonrandomized trial.

Table 3.1 Participant demographics and characteristics grouped by baseline BP (N=141).

	Baseline BP Category			
	All	Controlled	Stage 1	Stage 2
Participants, n	141	38	48	55
Age (y), mean (SD)	57.5 (13.9)	57.8 (16.0)	57.6 (12.6)	57.3 (13.5)
Female, n (%)	62 (44)	14 (37)	24 (50)	24 (44)
Weight (lb), mean (SD)	175.8 (48.4)	170.0 (41.6)	164.5 (52.3)	189.7 (45.7)
Baseline SBP (mmHg), mean (SD)	131.9 (11.5)	121.4 (6.1)	128.8 (7.1)	141.9 (9.3)
Baseline DBP (mmHg), mean (SD)	82.9 (9.0)	74.2 (4.4)	82.2 (6.4)	89.4 (8.0)
Taking hypertension medication, n (%)	118 (83.7)	32 (84)	39 (81)	47 (85)

a 20.3% drop rate from 12 weeks to 24 weeks. Reasons for participants withdrawing from the study included receiving new medical diagnoses (eg, cancer diagnosis), achieving a healthy BP, family emergencies, and other personal reasons. For the 141 participants who onboarded, Table 3.1 compares the characteristics between subgroups based on baseline BP classifications. The average age of participants was 57.5 (SD 13.9) years, and 44% (62/141) of the participants were female. For participants who had stage 2 hypertension at baseline, the average baseline BP was 141.9/89.4 mm Hg. In total, 83.7% (118/141) of the participants reported that they were taking antihypertensive medication at the beginning of the study.

As previously described, we asked participants each week to rate the difficulty of the recommendations they received on a scale from 1 to 5 and indicate whether they could follow each recommendation. This was done to assess compliance and the perceived difficulty of the

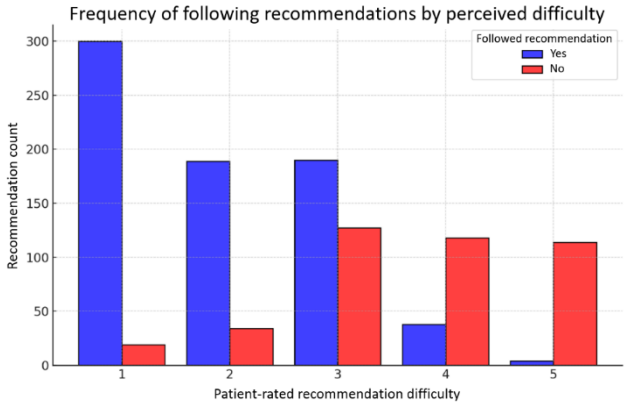

Figure 3.4 Histogram showing the number of recommendations adhered to based on their difficulty rating. The average difficulty rating for recommendations that were followed was 1.97, indicating lower difficulty, whereas the average for those not followed was 3.67, indicating higher difficulty.

recommendations. The histogram of difficulty ratings, divided into Yes and No responses, is shown in Figure 3.4. Recommendations were followed 63.64% (721/1133) of the time and not followed 36.36% (412/1133) of the time. The average difficulty rating for recommendations that were followed was 1.97, indicating lower difficulty, whereas the average for those not followed was 3.67, indicating higher difficulty. Evidently, there is a negative correlation between the perceived difficulty of a recommendation and its likelihood of being followed. We also tracked the number of unique recommendations each patient was sent. Out of the 37 unique recommendations, patients received an average of 9.4 (25%) unique recommendations each. The distribution of the number of unique recommendations is shown in Figure 3.5. The median (IQR) suggest a distribution close to normal. The maximum number of unique recommendations received by a single patient was as high as 21. These statistics demonstrate a broad range of recommendations given to the patients, covering various aspects of lifestyle.

An additional feasibility outcome we evaluated was participant experience as measured by responses to a study experience survey. As previously mentioned, this survey asked patients to rate the difficulty level of completing the study tasks, how useful they found the recommendations,

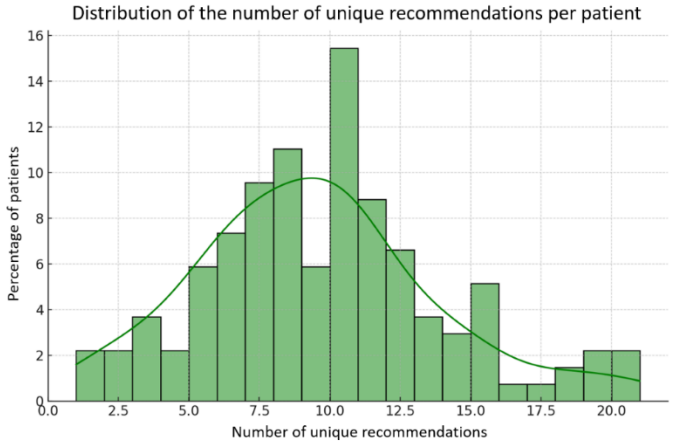

Figure 3.5 Distribution showing the number of unique recommendations sent to each patient. Patients received an average of 9.4 unique recommendations each.

and their experience using the app. In total, 70 participants responded to the survey. In total, 61% (43/70) of the participants responded that the study tasks were "easy" or "very easy" to incorporate into their daily routine, 51% (36/70) of the participants found the personalized recommendations to be "useful" or "very useful" compared to generic recommendations, and 86% (60/70) of the participants rated the app experience as "good" or "great."

3.3.2 BP Outcomes

For assessing BP outcomes, we used data from the 128 and 102 participants who completed 12 and 24 weeks in the study, respectively. Table 3.2 details the change in BP from baseline to 12 weeks. Across all participants, there was a statistically significant change of −5.6 mm Hg in SBP (t_{127}=7.6; P<.001; 95% CI −7.1 to −4.2) and −3.8 mm Hg in DBP (t_{127}=7.7; P<.001; 95% CI −4.7 to −2.8) after 12 weeks. Notably, 45.3% (58/128) of the participants achieved a clinically meaningful SBP drop of ≥5 mm Hg after 12 weeks. Table 3.3 details the change in BP from baseline to 24 weeks. For the participants who completed 24 weeks in the study, there was a statistically significant change of −8.1 mm Hg in SBP (t_{101}=8.1; P<.001; 95% CI −10.1 to −6.1)

Table 3.2 Comparison of average BP change at 12 weeks for different participant subgroups based on baseline BP (N=128).

	Participants, n	Change in BP at 12 weeks, Δmean (95% CI)	t test (*df*)	*P* value	≥ 5 mmHg reduction in SBP at 12 weeks, n (%)
SBP					
Overall	128	−5.6 (−7.1 to −4.2)	7.6 (127)	<.001	58 (45.3)
Controlled	31	−3.6 (−5.5 to −1.6)	3.7 (30)	.001	11 (35)
Stage 1	46	−2.6 (−4.8 to −0.5)	2.5 (45)	.02	14 (30)
Stage 2	51	−9.6 (−12.2 to −6.9)	7.3 (50)	<.001	33 (65)
DBP					
Overall	128	−3.8 (−4.7 to −2.8)	7.7 (127)	<.001	N/A
Controlled	31	−1.6 (−3.0 to −0.2)	2.3 (30)	.03	N/A
Stage 1	46	−3.1 (−4.4 to −1.7)	4.7 (45)	<.001	N/A
Stage 2	51	−5.7 (−7.6 to −3.9)	6.2 (50)	<.001	N/A

and −5.1 mm Hg in DBP (t_{101}=8.4; *P*<.001; 95% CI −6.2 to −3.9). In total, 58.8% (60/102) of the participants achieved a clinically meaningful SBP drop of ≥5 mm Hg after 24 weeks.

Participants with a baseline BP classified as stage 2 hypertension had the greatest change in BP and the greatest percentage of participants achieving a clinically meaningful SBP drop after 12 and 24 weeks. For these participants, SBP and DBP improved by −9.6 mm Hg (t_{50}=7.3; *P*<.001; 95% CI −12.2 to −6.9) and −5.7 mm Hg (t_{50}=6.2; *P*<.001; 95% CI −7.6 to −3.9) after 12 weeks, respectively, and −14.2 mm Hg (t_{36}=8.2; *P*<.001; 95% CI −17.7 to −10.7) and −8.1 mm Hg (t_{36}=7.0; *P*<.001; 95% CI −10.4 to −5.7) after 24 weeks, respectively. In total, 65% (33/51) and 78% (29/37) of the participants achieved a clinically meaningful SBP drop of ≥5 mm Hg after 12 and 24 weeks, respectively.

Table 3.3 Comparison of average BP change at 24 weeks for different participant subgroups based on baseline BP (N=102).

	Participants, n	Change in BP at 24 weeks, Δmean (95% CI)	t test (*df*)	*P* value	≥ 5 mmHg reduction in SBP at 24 weeks, n (%)
SBP					
Overall	102	−8.1 (−10.1 to −6.1)	8.1 (101)	<.001	60 (58.8)
Controlled	28	−3.9 (−7.1 to −0.8)	2.6 (27)	.02	14 (50)
Stage 1	37	−5.2 (−7.9 to −2.5)	3.9 (36)	<.001	17 (46)
Stage 2	37	−14.2 (−17.7 to −10.7)	8.2 (36)	<.001	29 (78)
DBP					
Overall	102	−5.1 (−6.2 to −3.9)	8.4 (101)	<.001	N/A
Controlled	28	−1.9 (−3.6 to −0.2)	2.3 (27)	.03	N/A
Stage 1	37	−4.4 (−6.0 to −2.8)	5.7 (36)	<.001	N/A
Stage 2	37	−8.1 (−10.4 to −5.7)	7.0 (36)	<.001	N/A

Another primary outcome we assessed was the percentage change of participants in different BP categories from baseline to 12 weeks and 24 weeks. Tables 3.4 and 3.5 detail this analysis. For participants completing 12 weeks in the study, the percentage of participants in the controlled range increased by 17.2% from 24.2% (31/128) to 41.4% (53/128; McNemar χ^2_1=3.0; *P*<.001). The percentage of participants with stage 2 hypertension decreased by 25% from 39.8% (51/128) to 14.8% (19/128; McNemar χ^2_1=4.0; *P*<.001) after 12 weeks. This means that 63% (32/51) of the patients with stage 2 hypertension at baseline moved into lower BP categories after 12 weeks. For those who completed 24 weeks in the study, the percentage in the controlled range increased by 26.5% from 27.5% (28/102) to 53.9% (55/102; McNemar χ^2_1=2.0; *P*<.001), and the stage 2 percentage decreased by 26.5% from 36.3% (37/102) to 9.8% (10/102; McNemar χ^2_1=3.0; *P*<.001). This means that 73% (27/37) of the patients with stage 2 hypertension at baseline moved

Table 3.4 Change in the percentage of participants in different BP categories from baseline to 12 weeks (N=128).

	Population at baseline, n (%)	Population at 12 weeks, n (%)	12-week difference, n (%)	McNeymar $\chi 2$ (*df*)	*P* value
Controlled	31 (24.2)	53 (41.4)	22 (17.2)	3.0 (1)	<.001
Stage 1	46 (35.9)	56 (43.8)	10 (7.8)	20.0 (1)	.20
Stage 2	51 (39.8)	19 (14.8)	−32 (−25)	4.0 (1)	<.001

into lower BP categories after 24 weeks. Note that the percentage changes for the stage 1 hypertension category from baseline to 12 weeks and 24 weeks were not statistically significant at the *P*=.05 level. The smaller change in the stage 1 hypertension population is due to a cascading effect where the number of participants moving from stage 2 into stage 1 was offset by the number of patients moving out of stage 1 and into the controlled BP category. For example, from baseline to 24 weeks, 18 participants moved from stage 2 to stage 1, and 17 participants moved from stage 1 to the controlled category.

3.3.3 Participant Engagement

We assessed participant engagement based on the percentage of active participants completing the program tasks each week. Figure 3.6 shows the weekly percentage of active patients measuring their BP, syncing their wearable device, and answering the questionnaire during the 24 weeks, respectively. We set an engagement goal of 90% for the study, which is represented by the red dashed lines in the figures. The average BP measurement engagement rate was 93% (SD 4.3%), and this rate was >90% for 19 (79%) out of 24 weeks. The average wearable syncing engagement rate was 94% (SD 2.4%), and this rate was >90% for 21 (88%) out of 24

Table 3.5 Change in the percentage of participants in different BP categories from baseline to 24 weeks (N=102).

	Population at baseline, n (%)	Population at 24 weeks, n (%)	24-week difference, n (%)	McNeymar $\chi 2$ (*df*)	*P* value
Controlled	28 (27.5)	55 (53.9)	27 (26.5)	2.0 (1)	<.001
Stage 1	37 (36.3)	37 (36.3)	0 (0)	N/A	N/A
Stage 2	37 (36.3)	10 (9.8)	−27 (−26.5)	3.0 (1)	<.001

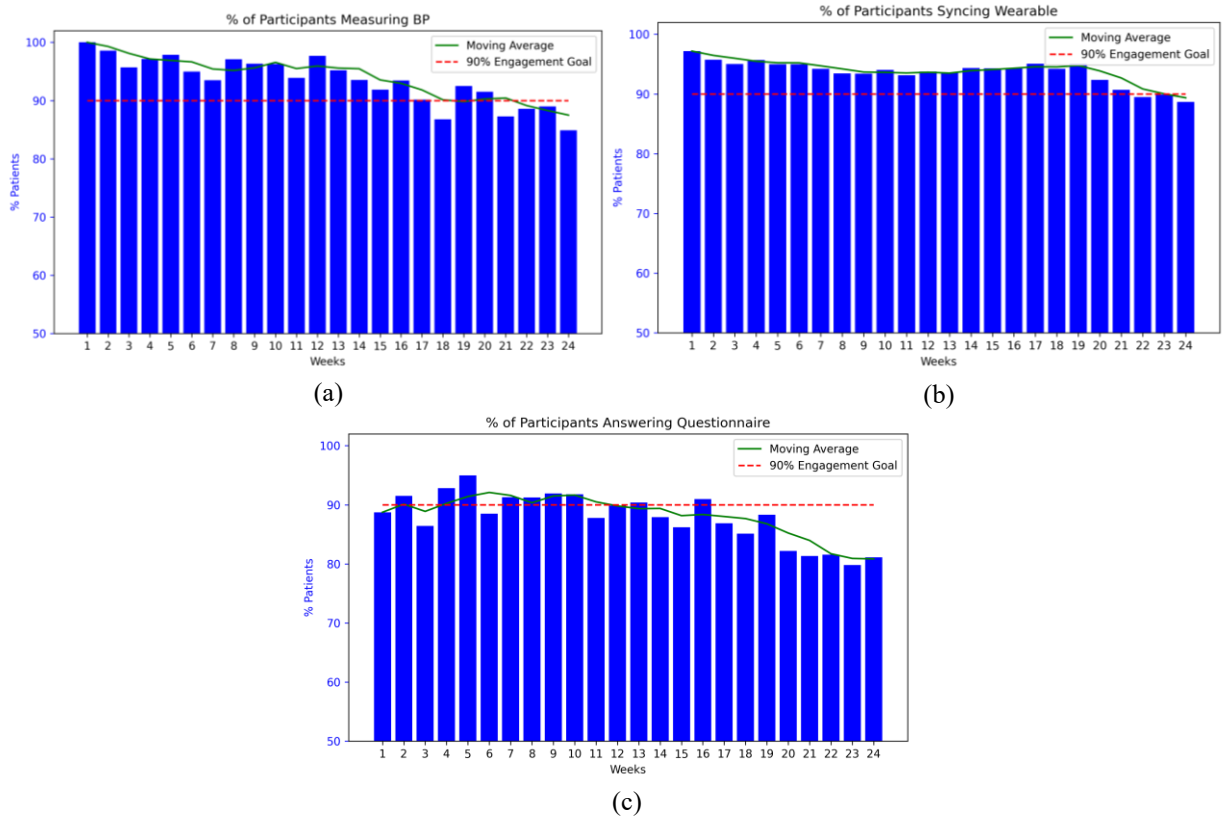

Figure 3.6 Percentage of active participants (a) measuring their BP (b) syncing their wearable device and (c) answering the mobile app questionnaire during the 24 weeks.

weeks. The average questionnaire engagement rate was 88% (SD 4.9%), and this rate was >90% for 10 (42%) out of 24 weeks.

Table 3.6 Participant escalations leading to manual care team outreaches for critical BP readings.

	Participants, n	Escalations, n	Participants escalated, n (%)
12 Weeks	128	8	5 (3.9)
24 Weeks	102	11	6 (5.9)

3.3.4 Clinician Intervention

Table 3.6 details the clinician intervention required during the program. For the 128 participants completing 12 weeks in the study, an escalation notification was sent to the care team 8 times. There were 3.9% (5/128) unique patients who required manual outreach during the first 12 weeks. For the 102 patients completing 24 weeks in the study, an escalation notification was sent to the PHSO care team 11 times. There were 5.9% (6/102) unique patients who required manual outreach during the 24 weeks.

3.4 Discussion

3.4.1 Principal Findings

This study aims to assess the effectiveness of a fully digital, autonomous, and AI-based lifestyle coaching program in achieving BP control and high engagement among adults with hypertension. The key components of this program included detailed lifestyle data collection via both wearables and questionnaires and weekly lifestyle recommendations based on personalized, AI-based analytics delivered via a mobile app. The guidance supported the participant's daily efforts to improve BP through behavioral changes that targeted physical activity, sleep hygiene, stress management, and dietary choices. Specifically, the program provided weekly guidance based on associations between lifestyle data and BP uncovered using ML and asked the participants to focus on the lifestyle factor with the greatest association. The precise lifestyle

recommendations enabled participants to focus on the most relevant aspect of their lifestyle as opposed to receiving general guidance. Our intervention approach aligns with the Fogg Behavioral Model, which states that 3 elements (ability, motivation, and prompts) are essential for behavior change [131]. By directing participants to focus on 1 lifestyle behavior at a time, the intervention simplified compliance and therefore increased the ability of the participants to adhere to the recommendations. This targeted strategy likely bolstered participants' motivation, as they could clearly see how specific lifestyle modifications directly influenced their BP. Each recommendation was delivered via a text message and prompted the user to take specific action. Furthermore, each recommendation was sent with a motivational message regarding their BP progress. We believe that this combination of personalized advice, ease of compliance, and motivational reinforcement contributed to our high engagement and improved BP outcomes.

We assessed multiple feasibility outcomes, including enrollment rate, adherence, and participant experience. In total, 59.9% (164/274) of the patients who initially expressed interest in joining the program ended up enrolling. Furthermore, although patients were recruited based on their last clinical BP reading, which required an SBP≥140 mm Hg or DBP≥90 mm Hg (stage 2 hypertension), many participants were not in the stage 2 range at baseline. Possible reasons for this include white coat hypertension [115] or that between the time of their last clinical BP reading and their enrollment in the study, they may have started taking BP medication or changed their diet. To improve the enrollment rate and ensure that patients who enroll have stage 2 hypertension, a new recruitment strategy is required. This new strategy could involve recruiting patients through PCP referrals. We hypothesize that this will increase the take-up rate due to increased trust from the more personal nature of the referral [122]. Furthermore, for the patients who are referred to the study, their PCPs would be instructed not to start the patients on any new BP medication or lifestyle

intervention before the study, except in critical cases. This would help ensure patients joining the study are indeed in the stage 2 hypertension category. Another feasibility outcome we assessed was participant experience. While most participants (43/70, 61%) found the study tasks easy to incorporate into their daily routine, a few (3/70, 4%) found it difficult. These included difficulty in measuring BP due to work schedules and travel, caregiving responsibilities, and equipment and syncing issues. To address these challenges, the intervention should be more context aware and adapt the program tasks and recommendations based on patients' circumstances. For example, a patient who works a night shift should not be asked to measure their BP at the same time or be given the same sleep recommendations as a patient who works during the day. Context-aware interventions would enhance the patient experience and increase the engagement rate.

Participants experienced a statistically significant decrease of 8.1 mm Hg and 5.1 mm Hg in SBP and DBP, respectively, after 24 weeks. Furthermore, this improvement was more pronounced in participants who started the program with stage 2 hypertension, achieving a 14.2 mm Hg and 8.1 mm Hg reduction in SBP and DBP, respectively. Reducing BP holds clinical significance not only for individuals with stage 2 hypertension but also for those with elevated BP or stage 1 hypertension. This is clinically meaningful as lower SBP values have been associated with progressively reduced risks of stroke, major cardiovascular events, and cardiovascular as well as all-cause mortalities [130]. In addition to BP improvement, the study demonstrates the intervention's ability to maintain sustained engagement. However, the engagement rate dropped during the last 4 weeks potentially because the participants whose BP had improved through the program may have reduced their engagement as they did not feel the urgent need. In this study, the participant tasks remain consistent; however, participants may find it useful if the requirements are adaptive based on their health condition and preferences. It is worthwhile to design a dynamic

mechanism that can adjust the extent and frequency of patient requirements based on the intervention progress. Both the BP and engagement results are achieved with minimal clinician intervention, primarily due to the autonomous nature of the intervention, demonstrating the potential scalability of this approach for hypertension management.

The observed BP improvement results from this study are comparable to those from clinician-led hypertension management programs [107-110]. The 3-month intervention program presented in the study by Wilson-Anumudu et al [107] combined lifestyle counseling with hypertension education, guided home BP monitoring, support for taking medications, and was led by either a registered nurse or certified diabetes care and education specialist. Patients with stage 2 hypertension who participated in this program experienced a 10.3 mm Hg and 6.5 mm Hg reduction in SBP and DBP, respectively, after 3 months. In the study by Milani et al [109], the 3-month digital intervention involved patients measuring their BP at least once per week and corresponding with pharmacists and health coaches to cocreate their treatment plan by choosing among various lifestyle modifications (eg, reducing dietary sodium) and medication options (eg, switching to generics or lower cost options). Patients with stage 2 hypertension participating in this program experienced a 14.0 mm Hg and 5.0 mm Hg reduction in SBP and DBP, respectively, after 3 months. Both interventions presented in the studies by Wilson-Anumudu et al [107] and Milani et al [109] assigned participants a designated hypertension coach who would provide lifestyle education and recommendations. These previous studies [107,109] primarily attribute their BP outcomes to the program's support led by health professionals who interpreted BP data and supported lifestyle change. While health coach–based programs can produce meaningful BP improvements, the reliance on health coaches is highly time and resource intensive. Consequently, these approaches have limited scalability and accessibility as an individual health coach can only

engage and care for a limited number of patients at a time. In contrast, our results demonstrate that a fully digital, AI-based lifestyle coaching program can produce clinically meaningful BP improvements comparable to those of programs led by health professionals. There is also potential for our approach to be used in conjunction with health coach–based programs. Under such a framework, our AI-based interactions and learnings from the patients can extend the reach of health coaches and provide them with more detailed insights about lifestyle factors impacting patients.

3.4.2 Study Limitations and Future Directions

As this was a single-arm nonrandomized study, it was not possible to conduct a causal analysis due to the lack of a control group. In addition, regression to the mean is another limitation as participants with initially high BP values may naturally converge toward the average over time. Therefore, to conduct causal analysis and account for regression to the mean, a randomized controlled trial may be conducted to draw stronger conclusions in a future study. To gain additional insights into the effectiveness of the program, we can randomize patients into different treatment arms by providing different versions of the program. This could include varying the frequency or content of the lifestyle recommendations across the different treatment arms. Furthermore, we could investigate which lifestyle interventions, for example, increasing steps or improving sleep hygiene, result in greater BP improvements. With careful design, we can create a multiarm trial to investigate optimal engagement strategies and recommendations for different types of patients. Another limitation of this study is selection bias as the participants self-selected to enroll after receiving the recruitment flyer. To address this, we plan to recruit patients through PCP referrals. PCPs will refer their patients with high cardiovascular risk, who can benefit from our intervention. As previously mentioned, we hypothesize that this will increase the take-up rate due to increased

trust from the more personal nature of the referral [122]. In addition, there is a need for a longer follow-up period as behavioral interventions can show improved outcomes during the first 6 months and then recidivism during the next 6 months. Finally, we did not collect socioeconomic data (eg, occupation, education, and income) from participants, preventing an analysis of how socioeconomic status impacts the program outcomes. In our future research, we will consider socioeconomic factors when analyzing the impact of the intervention. This analysis is imperative to ensure that the use of digital technologies does not contribute to an increased digital divide in health care and that all patients have equal access to high-quality health care [116,117].

3.4.3 Conclusions

To address the challenges of poor patient engagement due to generic, nonpersonalized lifestyle guidance and limited scalability of care due to human coaching models, we propose an AI-driven, autonomous, precise lifestyle coaching program for patients with hypertension. Patients who enrolled in the program experienced a significant improvement in BP. The program maintained a high engagement rate with minimal intervention from the care team. As the burden of hypertension increases globally, the necessity to develop new strategies to achieve hypertension control at scale is greater than ever. An AI-based, autonomous approach to hypertension-related lifestyle coaching can increase scalability and accessibility to effective BP management, ultimately improving the cardiovascular health of our community.

Chapter 3, in part, is from the material as it appears in the Journal of Medical Internet Research, Cardio, 2024, Leitner, Jared; Chiang, Po-Han; Agnihotri, Parag; Dey, Sujit. The dissertation author was the primary investigator and author of this paper.

CONCLUSION

In this dissertation, we investigate three applications of ML to real patient data to enable personalized, remote health monitoring and care delivery. In chapter 1, we demonstrate that we can use transfer learning to train personalized deep learning models for PPG-based BP estimation with limited patient data and achieve clinically meaningful accuracy. Furthermore, our approach enables continuous, noninvasive BP estimation, as opposed to the current standards of inflatable cuff-based measurement, which significantly limits the frequency of BP measurements, and the arterial catheter which is highly invasive and not practical for at-home measurement. In chapter 2, we propose a novel ML approach to estimate whether a patient has recovered from COVID-19 symptoms based on automatically collected wearable data. Our results demonstrate that ML-assisted remote monitoring using wearable data can supplement or replace manual daily COVID-19 symptom tracking, thereby enabling more optimal care of COVID-19 ambulatory patients at scale by remotely tracking patient recovery status. Finally, in chapter 3, we present the results of our trial with UCSD Health's Population Health Services Organization in which hypertension patients received an AI-driven, autonomous, precise lifestyle coaching intervention. Patients who enrolled experienced a significant improvement in BP and maintained a high engagement rate, while requiring minimal manual intervention from the care team. The outcomes demonstrate the potential of AI-based lifestyle coaching to increase scalability and accessibility to effective BP management.

In the future, we would like to extend our research in the following directions. Firstly, the training and inference for all ML models are implemented on a centralized server. For future work, we will investigate on-device model training either on a patient's cell phone or wearable device. This shift not only promises to enhance user engagement by achieving more real-time and

localized data analysis but also significantly mitigates risks associated with data transmission and enhances data privacy. Additionally, we plan to explore additional methods of data collection that are more patient friendly and context aware. For example, nutrition data remains a challenge for patients to collect. We will investigate computer vision and conversational AI-based approaches to nutrition data collection that fit more seamlessly into the daily routines of patients. Furthermore, to conduct causal analysis for our AI-based lifestyle interventions, a randomized controlled trial can be conducted to draw stronger conclusions in a future study. This future study can examine how different subgroups, based on demographic, socioeconomic, and contextual factors, are impacted by the AI-based intervention to ensure patients of diverse backgrounds can benefit from healthcare technology advancements. Finally, leveraging large language models (LLMs) could be a transformative aspect of our research. LLMs can be utilized to generate personalized patient communication and support, analyze large volumes of unstructured data for insights, or even simulate patient interactions for better model training without compromising privacy. These applications of LLMs represent a frontier in medical AI research that holds potential for significant contributions to personalized and accessible healthcare solutions.

REFERENCES

[1] Fryar, C. D., Ostchega, Y., Hales, C. M., Zhang, G., & Kruszon-Moran, D. (2017). Hypertension Prevalence and Control Among Adults: United States, 2015-2016. *NCHS data brief*, (289), 1–8.

[2] Centers for Disease Control and Prevention, National Center for Health Statistics. Underlying cause of death 1999–2013 on CDC WONDER online database, released 2015.

[3] The World Health Organization, "A Global Brief on Hypertension," April 2013.

[4] American College of Cardiology, "Guideline for the Prevention, Detection, Evaluation, and Management of High Blood Pressure in Adults," Journal of the American College of Cardiology, pp.4-28, 2017.

[5] Ogedegbe, G., & Pickering, T. (2010). Principles and techniques of blood pressure measurement. *Cardiology clinics, 28*(4), 571–586. https://doi.org/10.1016/j.ccl.2010.07.006

[6] Cho, J. (2019). Current Status and Prospects of Health-Related Sensing Technology in Wearable Devices. *Journal of Healthcare Engineering, 2019*, 1–8. https://doi.org/10.1155/2019/3924508

[7] Liang, Y., Chen, Z., Ward, R., & Elgendi, M. (2018). Photoplethysmography and Deep Learning: Enhancing Hypertension Risk Stratification. *Biosensors, 8*(4), 101. https://doi.org/10.3390/bios8040101

[8] Hasanzadeh, N., Ahmadi, M. M., & Mohammadzade, H. (2020). Blood Pressure Estimation Using Photoplethysmogram Signal and Its Morphological Features. *IEEE Sensors Journal, 20*(8), 4300–4310. https://doi.org/10.1109/jsen.2019.2961411

[9] Radha, M., de Groot, K., Rajani, N., Wong, C. C. P., Kobold, N., Vos, V., Fonseca, P., Mastellos, N., Wark, P. A., Velthoven, N., Haakma, R., & Aarts, R. M. (2019). Estimating blood pressure trends and the nocturnal dip from photoplethysmography. *Physiological Measurement, 40*(2), 025006. https://doi.org/10.1088/1361-6579/ab030e

[10] Slapničar, G., Luštrek, M., & Marinko, M. (2018). Continuous Blood Pressure Estimation from PPG Signal. *Informatica (Lithuanian Academy of Sciences), 42*(1).

[11] Khalid, S. G., Zhang, J., Chen, F., & Zheng, D. (2018). Blood Pressure Estimation Using Photoplethysmography Only: Comparison between Different Machine Learning Approaches. *Journal of Healthcare Engineering, 2018*, 1–13. https://doi.org/10.1155/2018/1548647

[12] Lim, P. K., Ng, S.-C., Lovell, N. H., Yu, Y. P., Tan, M. P., McCombie, D., Lim, E., & Redmond, S. J. (2018). Adaptive template matching of photoplethysmogram pulses to detect motion artefact. *Physiological Measurement, 39*(10), 105005. https://doi.org/10.1088/1361-6579/aadf1e

[13] Leitner, J., Chiang, P.-H., & Dey, S. (2019). Personalized Blood Pressure Estimation using Photoplethysmography and Wavelet Decomposition. *2019 IEEE International Conference on*

E-Health Networking, Application & Services (HealthCom). https://doi.org/10.1109/healthcom46333.2019.9009587

[14] Xia, P., Hu, J., & Peng, Y. (2017). EMG-Based Estimation of Limb Movement Using Deep Learning With Recurrent Convolutional Neural Networks. *Artificial Organs, 42*(5), E67–E77. https://doi.org/10.1111/aor.13004

[15] Luay Fraiwan, & Khaldon Lweesy. (2017). Neonatal sleep state identification using deep learning autoencoders. *2017 IEEE 13th International Colloquium on Signal Processing & its Applications (CSPA)*. https://doi.org/10.1109/cspa.2017.8064956

[16] Acharya, U. R., Fujita, H., Lih, O. S., Hagiwara, Y., Tan, J. H., & Adam, M. (2017). Automated detection of arrhythmias using different intervals of tachycardia ECG segments with convolutional neural network. *Information Sciences*, *405*, 81–90. https://doi.org/10.1016/j.ins.2017.04.012

[17] Slapničar, G., Mlakar, N., & Luštrek, M. (2019). Blood Pressure Estimation from Photoplethysmogram Using a Spectro-Temporal Deep Neural Network. *Sensors, 19*(15), 3420. https://doi.org/10.3390/s19153420.

[18] Schlesinger, O., Nitai Vigderhouse, Eytan, D., & Moshe, Y. (2020). Blood Pressure Estimation From PPG Signals Using Convolutional Neural Networks And Siamese Network. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. https://doi.org/10.1109/icassp40776.2020.9053446

[19] Wang, C., Yang, F., Yuan, X., Zhang, Y., Chang, K., & Li, Z. (2020). An end-to-end neural network model for blood pressure estimation using PPG signal. *Artificial Intelligence in China*. https://doi.org/10.1007/978-981-15-0187-6_30

[20] Pan, S. J., & Yang, Q. (2010). A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, *22*(10), 1345–1359. https://doi.org/10.1109/tkde.2009.191

[21] Sainath, T. N., Vinyals, O., Senior, A., & Sak, H. (2015). Convolutional, Long Short-Term Memory, fully connected Deep Neural Networks. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. https://doi.org/10.1109/ICASSP.2015.7178838

[22] Johnson, A., Pollard, T., Shen, L., Lehman, L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L., & Mark, R. (2016). MIMIC-III, a freely accessible critical care database. *Sci Data* 3, 160035. https://doi.org/10.1038/sdata.2016.35

[23] Goldberger, A. L., Amaral, L. A. N., Glass, L., Hausdorff, J. M., Ivanov, P. Ch., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C.-K., & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet. *Circulation*, *101*(23). https://doi.org/10.1161/01.cir.101.23.e215

[24] Johnson, A., Pollard, T., & Mark, R. (2016). MIMIC-III Clinical Database (version 1.4). *PhysioNet*. https://doi.org/10.13026/C2XW26

[25] Choi, A., & Shin, H. (2017). Photoplethysmography sampling frequency: pilot assessment of how low can we go to analyze pulse rate variability with reliability? *Physiological Measurement*, *38*(3), 586–600. https://doi.org/10.1088/1361-6579/aa5efa

[26] Glorot, X. & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. *Proc. AISTATS*.

[27] Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. https://doi.org/10.48550/arXiv.1412.3555

[28] Ioffe, S. & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. https://doi.org/10.48550/arXiv.1502.03167

[29] Chen, H., Lundberg, S, Erion, G., Kim, J. H., & Lee, S. (2020). Deep Transfer Learning for Physiological Signals.

[30] Zhang, Y. & Yang, Q. (2017). A survey on multi-task learning. https://doi.org/10.48550/arXiv.1707.08114

[31] Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., & Lerer, A. (2017). Automatic differentiation in PyTorch.

[32] D. Eigen, J. Rolfe, R. Fergus and Y. LeCun, "Understanding deep architectures using a recursive convolutional network", *arXiv:1312.1847*, 2013.

[33] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization", *Proc. Int. Conf. Learn. Represent*., pp. 1-41, 2015.

[34] L. Prechelt, "Early stopping—But when?" in Neural Networks: Tricks of the Trade, Springer, pp. 53-67, 2012.

[35] E. O'Brien, B. Waeber, G. Parati, J. Stassen and M. Myers, "Blood pressure measuring devices: recommendations of the European Society of Hypertension", vol. 322, no. 7285, pp. 531-536, 2001.

[36] "Non-invasive sphygmomanometers—Part 2: Clinical investigation of automated measurement type" in , Arlington, VA, USA, 2016.

[37] Association for the Advancement of Medical Instrumentation (AAMI), "American National Standard Manual," *Electronic or Automated Sphygmonanometers*, AASI/AAMI SP 10:2002, 2003.

[38] D. G. Altman and J. M. Bland, "Measurement in medicine: The analysis of method comparison studies", *Statistician*, vol. 32, pp. 307-317, 1983.

[39] J. Benesty, J. Chen, Y. Huang and I. Cohen, "Pearson Correlation Coefficient" in Noise Reduction in Speech Processing, Springer, pp. 1-4, 2009.

[40] "WHO Coronavirus Disease (COVID-19) Dashboard." World Health Organization, covid19.who.int/

[41] "CDC COVID Data Tracker." Centers for Disease Control and Prevention, covid.cdc.gov/covid-data-tracker/.

[42] "COVIDView: A Weekly Surveillance Summary of U.S. COVID-19 Activity." Centers for Disease Control and Prevention, www.cdc.gov/coronavirus/2019-ncov/covid-data/covidview/index.html.

[43] D. M. Roblyer, "Perspective on the increasing role of optical wearables and remote patient monitoring in the COVID-19 era and beyond," in *Journal of Biomedical Optics* 25(10) 102703 (21 October 2020) https://doi.org/10.1117/1.JBO.25.10.102703

[44] A. Mahajan, G. Pottie, and W. Kaiser, "Transformation in Healthcare by Wearable Devices for Diagnostics and Guidance of Treatment," in *ACM Trans. Comput. Healthcare* 1, 1, Article 2 (February 2020), 12 pages. DOI: https://doi.org/10.1145/3361561

[45] M. M. Islam, S. Mahmud, L. J. Muhammad, M. R. Islam, S. Nooruddin and S. I. Ayon, "Wearable technology to assist the patients infected with novel coronavirus (COVID-19)", *Social Netw. Comput. Sci*., vol. 1, no. 6, pp. 320, Nov. 2020.

[46] N. Ji, T. Xiang, P. Bonato, N. H. Lovell, S. Y. Ooi, D. A. Clifton, M. Akay, X. R. Ding, B. P. Yan, V. Mok, D. I. Fotiadis and Y. T. Zhang, "Recommendation to Use Wearable-based mHealth in Closed-Loop Management of Acute Cardiovascular Disease Patients during the COVID-19 Pandemic," in *IEEE Journal of Biomedical and Health Informatics*, doi: 10.1109/JBHI.2021.3059883.

[47] M. Au-Yong-Oliveira, A. Pesqueira, M. J. Sousa, F. D. Mas, and M. Soliman, "The Potential of Big Data Research in HealthCare for Medical Doctors' Learning," in *Journal of Medical Systems* 45, 13 (2021). https://doi.org/10.1007/s10916-020-01691-7

[48] C. Menni, A. M. Valdes, M.B. Freidin, C. H. Sudre, L. H. Nguyen, D. A. Drew, S. Ganesh, T. Varsavsky, M. J. Cardoso, J. S. El-Sayed Msoustafa, A. Visconti, P. Hysi, R. C. E. Bowyer, M. Mangino, M. Falchi, J. Wolf, S. Ourselin, A. T. Chan, C. J. Steves and T. D. Spector, "Real-time tracking of self-reported symptoms to predict potential COVID-19," *Nat Med*, vol. 26, pp. 1037-1040, 2020

[49] M. Zens, A, Brammertz, J. Herpich, N. Sudkamp, and M. Hinterseer, "App-based tracking of self-reported covid-19 symptoms: Analysis of questionnaire data," *Journal of Medical Internet Research*, 22(9):e21956, 2020.

[50] Y. Zoabi, S. Deri-Rozov, and N. Shomron, "Machine learning-based prediction of covid-19 diagnosis based on symptoms," *npj digital medicine*, vol. 4, no. 1, pp. 1–5, 2021.

[51] T. Mishra, M. Wang, A. A. Metwally, G. K. Bogu, A. W. Brooks, A. Bahmani, A. Alavi, A. Celli, E. Higgs, O. Dagan-Rosenfeld, B. Fay, S. Kirkpatrick, R. Kellogg, M. Gibson, T. Wang, E. M. Hunting, P. Mamic, A. B. Ganz, B. Rolnik, X. Li and M. P. Snyder, "Pre-

symptomatic detection of COVID-19 from smartwatch data", *Nat. Biomed. Eng*., vol. 4, no. 12, pp. 1208-1220, 2020.

[52] A. Alavi, G. K. Bogu, M. Wang, E. S. Rangan, A. W. Brooks, Q. Wang, E. Higgs, A. Celli, T. Mishra, A. A. Metwally, K. Cha, P. Knowles, A. A. Alavi, R. Bhasin, S. Panchamukhi, D. Celis, T. Aditya, A. Honkala, B. Rolnik, E. Hunting, O. Dagan-Rosenfeld, A. Chauhan, J. W. Li, C. Bejikian, V. Krishnan, L. McGuire, X. Li, A. Bahmani and M. P. Snyder, Real-time alerting system for COVID-19 and other stress events using wearable data. Nat Med (2021). https://doi.org/10.1038/s41591-021-01593-2

[53] G. Quer, J. M. Radin, M. Gadaleta, K. Baca-Motes, L. Ariniello, E. Ramos, V. Kheterpal, E. J. Topol and S. R. Steinhubl, "Wearable sensor data and self-reported symptoms for COVID-19 detection," *Nat Med* (2020). https://doi.org/10.1038/s41591-020-1123-xs

[54] A. Natarajan, H.W. Su, and C. Heneghan, "Assessment of physiological signs associated with COVID-19 measured using wearable devices," *npj Digit. Med.* 3, 156 (2020). https://doi.org/10.1038/s41746-020-00363-7

[55] M. Gadaleta, J. M. Radin, K. Baca-Motes, E. Ramos, V. Kheterpal, E. J. Topol, S. R. Steinhubl and G. Quer, Passive detection of COVID-19 with wearable sensors and explainable machine learning algorithms. npj Digit. Med. 4, 166 (2021). https://doi.org/10.1038/s41746-021-00533-1

[56] A. Salama, A. Darwsih, and A.E. Hassanien, "Artificial Intelligence Approach to Predict the COVID-19 Patient's Recovery," *Digital Transformation and Emerging Technologies for Fighting COVID-19 Pandemic: Innovative Approaches*, vol. 322, pp. 121-133, 2021.

[57] L.J. Muhammad, M.M. Islam, S.S. Usman, and S.I. Ayon, "Predictive Data Mining Models for Novel Coronavirus (COVID-19) Infected Patients' Recovery," *SN COMPUT. SCI.* 1, 206 (2020).

[58] Channa, A., Popescu, N., Skibinska, J., & Burget, R. (2021). The rise of wearable devices during the COVID-19 pandemic: A systematic review. Sensors, 21(17), 5787.

[59] Garmin, and Garmin Ltd. or its subsidiaries. "Garmin Vivosmart® 4: Fitness Activity Tracker: Pulse Ox." Garmin, buy.garmin.com/en-US/US/p/605739.

[60] S. Shah, K. Majmudar, A. Stein, N. Gupta, S. Suppes, M. Karamanis, J. Capannari, S. Sethi and C. Patte, "Novel use of home pulse oximetry monitoring in COVID-19 patients discharged from the emergency department identifies need for hospitalization," *Acad Emerg Med* 2020; 27. doi: https://doi.org/10.1111/acem.10453

[61] "Overview: Health API: Garmin Developers." Overview | Health API | Garmin Developers, developer.garmin.com/health-api/overview/.

[62] "When You Can Be Around Others After You Had or Likely Had COVID-19." *Centers for Disease Control and Prevention*, www.cdc.gov/coronavirus/2019-ncov/if-you-are-sick/end-home-isolation.html.

[63] "Post-COVID Conditions." *Centers for Disease Control and Prevention*, www.cdc.gov/coronavirus/2019-ncov/long-term-effects.html.

[64] S. Ikegami, R. Benirschke, T. Flanagan, N. Tanna, T. Klein, R. Elue, P. Debosz, J. Mallek, G. Wright, P. Guariglia, J. Kang and T. J. Gniadek, "Persistence of SARS-CoV-2 nasopharyngeal swab PCR positivity in COVID-19 convalescent plasma donors," *Transfusion.* 2020; 60: 2962– 2968. https://doi.org/10.1111/trf.16015

[65] Spearman Rank Correlation Coefficient. In: The Concise Encyclopedia of Statistics. Springer, New York, NY. https://doi.org/10.1007/978-0-387-32833-1_379

[66] J. Cline. "Flu Season and Sleep." *Psychology Today*, Sussex Publishers, 31 Dec. 2014.

[67] L. Breiman, "Random Forest," *Machine Learning*, vol. 45, no.1, pp. 5–32, 2001

[68] B. Efron and R. Tibshirani, "An introduction to the bootstrap," *CRC Press*, 1994.

[69] V. F. Rodriguez-Galiano, B. Ghimire, J. Rogan, M. Chica-Olmo, and J. P. Rigol-Sanchez, "An assessment of the effectiveness of a random forest classifier for land-cover classification," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 67, pp. 93–104, 2012.

[70] P. Chiang and S. Dey, "Personalized Effect of Health Behavior on Blood Pressure: Machine Learning Based Prediction and Recommendation," in *Proc. of IEEE International Conference on E-health Networking, Application & Services (Healthcom'18),* Ostrava, Czech, 2018

[71] J. Leitner, P. Chiang and S. Dey, "Personalized Blood Pressure Estimation Using Photoplethysmography: A Transfer Learning Approach," in *IEEE Journal of Biomedical and Health Informatics*, doi: 10.1109/JBHI.2021.3085526.

[72] C. L. Stewart, A. Folarin and R. Dobson, "Personalized acute stress classification from physiological signals with neural processes", 2020.

[73] D. Lopez-Martinez and R. Picard, "Multi-task neural networks for personalized pain recognition from physiological signals," *2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, San Antonio, TX, 2017, pp. 181-184, doi: 10.1109/ACIIW.2017.8272611.

[74] P. Chiang and S. Dey, "Offline and Online Learning Techniques for Personalized Blood Pressure Prediction and Health Behavior Recommendations," in *IEEE Access*, vol. 7, pp. 130854-130864, 2019, doi: 10.1109/ACCESS.2019.2939218.

[75] M. Grandini, E. Bagli, and G. Visani, "Metrics for multi-class classification: an overview," arXiv preprint arXiv:2008.05756, 2020.

[76] R. E. Wright, "Logistic regression," in L. G. Grimm & P. R. Yarnold (Eds.), *Reading and understanding multivariate statistics* (p. 217–244). American Psychological Association.

[77] A. Mucherino, P.J. Papajorgji, and P.M. Pardalos, "K-nearest neighbor classification," *Data mining in agriculture*. Springer, New York, NY, 2009. 83-106.

[78] J. A. K. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural Process. Lett.*, vol. 9, no. 3, pp. 293–300, 1999.

[79] D. Rumelhart, G. Hinton, and R. Williams, "Learning representations by back-propagating errors," *Cogn. Model*., vol. 5, no. 3, p. 1, 1988.

[80] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, pp. 1735-1780, 1997.

[81] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014

[82] S. Saeb, L. Lonini, A. Jayaraman, D. C. Mohr and K. P. Kording, "The need to approximate the use-case in clinical machine learning," *Gigascience* 2017;6(5):1–9.

[83] L. Shapley, "A value for n-person games." *Contributions to the Theory of Games*, vol. 2, no. 28, pp. 307-317, 1953.

[84] S. Cohen, E. Ruppin and G. Dror, "Feature Selection Based on the Shapley Value," in *International Joint Conferences on Artificial Intelligence*, vol. 5, pp. 665-670, 2005.

[85] S. Lundberg and S. Lee, "A unified approach to interpreting model predictions," in *Advances in neural information processing systems*, pp. 4765-4774, 2017.

[86] S. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal and S. I. Lee, "From local explanations to global understanding with explainable AI for trees," *Nat Mach Intell* 2, pp. 56-67, 2020.

[87] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems?" *The journal of machine learning research*, vol. 15, no. 1 pp. 3133-3181, 2014.

[88] S. Wang, C. Aggarwal, and H. Liu. "Using a random forest to inspire a neural network and improving on it," In Proceedings of the *SIAM international conference on data mining*, Houston, TX, USA, 2017.

[89] A. Ramsetty and C. Adams, "Impact of the digital divide in the age of COVID-19," *Journal of the American Medical Informatics Association*, vol. 27, pp. 1147-1148, 2020.

[90] Fryar C, Ostchega Y, Hales C, Zhang G, and Kruszon-Moran D, Hypertension prevalence and control among adults: United States, 2015–2016, NCHS data brief, no. 289, 2017.

[91] P. K. Whelton, R. M. Carey, W. S. Aronow, D. E. Casey Jr, K. J. Collins, C. D. Himmelfarb, S. M. DePalma, S. Gidding, K. A. Jamerson, D. W. Jones, E. J. MacLaughlin, P. Muntner, B. Ovbiagele, S. C. Smith Jr, C. C. Spencer, R. S. Stafford, S. J. Taler, R. J. Thomas,

K. A. Williams Sr, J. D. Williamson and J. T. Wright Jr, Prevention, Detection, Evaluation, and Management of High Blood Pressure in Adults: Executive Summary: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines, Journal of the American College of Cardiology, vol. 71, no. 19, pp. 2273-2275, 2018.

[92] Centers for Disease Control and Prevention. Hypertension Cascade: Hypertension Prevalence, Treatment and Control Estimates Among U.S. Adults Aged 18 Years and Older Applying the Criteria from the American College of Cardiology and American Heart Association's 2017 Hypertension Guideline. NHANES 2015–2018.

[93] Lebeau J, Cadwallader J, Aubin-Auger I, Mercier A, Pasquet T, Rusch E, Hendrickx K, and Vermeire E. The concept and definition of therapeutic inertia in hypertension in primary care: a qualitative systematic review. BMC Fam Pract 2014 Jul 02; 15:130.

[94] Kirkland EB, Heincelman M, Bishu KG, Schumann SO, Schreiner A, Axon RN, Mauldin PD, and Moran WP. Trends in healthcare expenditures among US adults with hypertension: national estimates, 2003–2014. J Am Heart Assoc. 2018.

[95] Blood Pressure Lowering Treatment Trialists' Collaboration. Pharmacological blood pressure lowering for primary and secondary prevention of cardiovascular disease across different levels of blood pressure: an individual participant-level data meta-analysis. Lancet 2021 May 01;397(10285):1625-1636.

[96] Shimbo D, Artinian NT, Basile JN, Krakoff LR, Margolis KL, Rakotz MK, American Heart Association the American Medical Association. Self-measured blood pressure monitoring at home: a joint policy statement from the American Heart Association and American Medical Association. Circulation 2020 Jul 28;142(4): e42-e63.

[97] Muntner P, Shimbo D, Carey RM, Charleston JB, Gaillard T, Misra S, Myers MG, Ogedegbe G, Schwartz JE, Townsend RR, Urbina EM, Viera AJ, White WB, and Wright JT. Measurement of blood pressure in humans: a scientific statement from the American Heart Association. Hypertension. 2019;73(5): e35-e66.

[98] Tucker KL, Sheppard JP, Stevens R, Bosworth HB, Bove A, Bray EP, Earle K, George J, Godwin M, Green BB, Hebert P, Hobbs Rs, Kantola I, Kerry SM, Leiva A, Magid DJ, Mant J, Margolis KL, McKinstry B, McLaughlin MA, Omboni S, Ogedegbe O, Parati G, Qamar N, Tabaei BP, Varis J, Verberk WJ, Wakefield BJ, and McManus RJ. Self-monitoring of blood pressure in hypertension: A systematic review and individual patient data meta-analysis. PLoS Med 2017 Sep;14(9): e1002389.

[99] Uhlig K, Patel K, Ip S, Kitsios GD, Balk EM. Self-measured blood pressure monitoring in the management of hypertension: a systematic review and meta-analysis. Ann Intern Med. 2013;159(3):185-194.

[100] Appel L, Champagne C, Harsha D, Cooper L, Obarzanek E, Elmer P, Stevens VJ, Vollmer WM, Lin PH, Svetkey LP, Stedman SW, and Young DR. Effects of comprehensive lifestyle

modification on blood pressure control: main results of the PREMIER clinical trial, Journal of the American Medical Association, vol 289, pp. 2083–2093, 2003.

[101] Doughty K, Del Pilar N, Audette A, Katz D, Lifestyle Medicine and the Management of Cardiovascular Disease, Current cardiology reports, vol. 19, no. 11, pp. 116, 2017.

[102] Covassin N and Singh P, Sleep duration and cardiovascular disease risk: epidemiologic and experimental evidence, Sleep medicine clinics. vol. 11, no. 1, pp. 81-89, 2016.

[103] Cornelissen V and Smart N, Exercise Training for Blood Pressure: A Systematic Review and Meta-analysis, Journal of the American Heart Association, 2013.

[104] Eckel RH, Jakicic JM, Ard JD, de Jesus JM, Miller NH, Hubbard VS, Lee IM, Lichtenstein AH, Loria CM, Millen BE, Nonas CA, Sacks FM, Smith SC, Svetkey LP, Wadden TA, and Yanovski SZ. 2013 AHA/ACC Guideline on Lifestyle Management to Reduce Cardiovascular Risk: A Report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines, Circulation (2014).

[105] Bakris G, Ali W, and Parati G. ACC/AHA Versus ESC/ESH on Hypertension Guidelines, J Am Coll Cardiol. 2019 Jun, 73 (23) 3018–3026.

[106] Bell RA and Kravitz RL, Physician counseling for hypertension: What do doctors really do? Patient Educ Couns., 72 (2008), pp. 115-121

[107] Wilson-Anumudu F, Quan R, Cerrada C, Juusola J, Castro Sweet C, Bradner Jasik C, Turken M, Pilot Results of a Digital Hypertension Self-Management Program Among Adults with Excess Body Weight: Single-Arm Nonrandomized Trial, JMIR Form Res. 2022 Mar 30;6(3): e33057. doi: 10.2196/33057. PMID: 35353040; PMCID: PMC9008519.

[108] Toro-Ramos T, Kim Y, Wood M, Rajda J, Niejadlik K, Honcz J, Marrero D, Fawer A, and Michaelides A. Efficacy of a mobile hypertension prevention delivery platform with human coaching. J Hum Hypertens 31, 795–800 (2017).

[109] Milani RV, Lavie CJ, Bober RM, Milani AR, Ventura HO. Improving Hypertension Control and Patient Engagement Using Digital Tools. Am J Med 2017 Jan;130(1):14-20.

[110] Mao AY, Chen C, Magana C, Caballero Barajas K, Olayiwola JN, A Mobile Phone-Based Health Coaching Intervention for Weight Loss and Blood Pressure Reduction in a National Payer Population: A Retrospective Study, JMIR Mhealth Uhealth 2017;5(6):e80

[111] Branch OH, Rikhy M, Auster-Gussman LA, Lockwood KG, Graham SA, Relationships Between Blood Pressure Reduction, Weight Loss, and Engagement in a Digital App-Based Hypertension Care Program: Observational Study, JMIR Form Res. 2022 Oct 27;6(10):e38215. doi: 10.2196/38215. PMID: 36301618; PMCID: PMC9650575.

[112] Leitner J, Chiang P, Khan B, Dey S, An mHealth Lifestyle Intervention Service for Improving Blood Pressure using Machine Learning and IoMTs, 2022 IEEE International

Conference on Digital Health (ICDH), 2022, pp. 142-150, doi: 10.1109/ICDH55609.2022.00030.

[113] Monitoring Your Blood Pressure at Home. Heart Attack and Stroke Symptoms. URL: https://www.heart.org/en/health-topics/high-blood-pressure/understanding-blood-pressure-readings/monitoring-your-blood-pressure-at-home

[114] Alagappan R and Das S, Uncovering Twilio: Insights into Cloud Communication Services.

[115] Franklin SS, Thijs L, Hansen TW, O'Brien E, Staessen JA. White-coat hypertension: new insights from recent studies. Hypertension. 2013; 62:982–987.

[116] Ramsetty A and Adams C, Impact of the digital divide in the age of COVID-19, J. Amer. Med. Inform. Assoc., vol. 27, pp. 1147–1148, 2020.

[117] Rodriguez JA, Clark CR, Bates DW. Digital Health Equity as a Necessity in the 21st Century Cures Act Era. JAMA. 2020;323(23):2381–2382.

[118] Siervo M, Lara J, Chowdhury S, Ashor A, Oggioni C, Mathers JC. Effects of the Dietary Approach to Stop Hypertension (DASH) diet on cardiovascular risk factors: a systematic review and meta-analysis. British Journal of Nutrition. 2015;113(1):1-15.

[119] Henriksen A, Haugen Mikalsen M, Woldaregay A, Muzny M, Hartvigsen G, Hopstock L, Grimsgaard S. Using Fitness Trackers and Smartwatches to Measure Physical Activity in Research: Analysis of Consumer Wrist-Worn Wearables. J Med Internet Res 2018;20(3):e110

[120] Germini F, Noronha N, Borg Debono V, Abraham Philip B, Pete D, Navarro T, Keepanasseril A, Parpia S, de Wit K, Iorio A. Accuracy and Acceptability of Wrist-Wearable Activity-Tracking Devices: Systematic Review of the Literature. J Med Internet Res 2022;24(1):e30791

[121] Miller DJ, Sargent C, Roach GD. A Validation of Six Wearable Devices for Estimating Sleep, Heart Rate and Heart Rate Variability in Healthy Adults. Sensors. 2022; 22(16):6317.

[122] Pfaff E, Lee A, Bradford R, Pae J, Potter C, Blue P, Knoepp P, Thompson K, Roumie CL, Crenshaw D, Servis R, and DeWalt DA. Recruiting for a pragmatic trial using the electronic health record and patient portal: Successes and lessons learned. J Am Med Inform Assoc 2019 Jan 01;26(1):44-49

[123] Yuenyongchaiwat K, Pipatsitipong D, Sangprasert P. Increasing walking steps daily can reduce blood pressure and diabetes in overweight participants. Diabetol Int 9, 75–79 (2018).

[124] Lefferts E, Saavedra J, Song B, Brellenthin A, Pescatello L, Lee D. Increasing Lifestyle Walking by 3000 Steps per Day Reduces Blood Pressure in Sedentary Older Adults with Hypertension: Results from an e-Health Pilot Study. Journal of Cardiovascular Development and Disease. 2023; 10(8):317.

[125] Bock J, Vungarala S, Covassin N, Somers V. Sleep Duration and Hypertension: Epidemiological Evidence and Underlying Mechanisms, American Journal of Hypertension, Volume 35, Issue 1, January 2022, Pages 3–11.

[126] Ali W, Gao G, Bakris G. Improved Sleep Quality Improves Blood Pressure Control among Patients with Chronic Kidney Disease: A Pilot Study. Am J Nephrol 11 March 2020; 51 (3): 249–254.

[127] Ponte Márquez PH, Feliu-Soler A, Solé-Villa MJ, Matas-Pericas L, Filella-Agullo D, Ruiz-Herrerias M, Soler-Ribaudi J, Roca-Cusachs Coll A, and Arroyo-Díaz JA. Benefits of mindfulness meditation in reducing blood pressure and stress in patients with arterial hypertension. J Hum Hypertens 33, 237–247 (2019).

[128] Gupta DK, Lewis CE, Varady KA, Su YR, Madhur MS, Lackland DT, Reis JP, Wang TJ, Lloyd-Jones DM, and Allen NB. Effect of Dietary Sodium on Blood Pressure: A Crossover Trial. JAMA. 2023;330(23):2258–2266. doi:10.1001/jama.2023.23651

[129] He FJ, Tan M, Ma Y, and MacGregor GA. Salt Reduction to Prevent Hypertension and Cardiovascular Disease: JACC State-of-the-Art Review. J Am Coll Cardiol. 2020 Feb, 75 (6) 632–647.

[130] Bundy JD, Li C, Stuchlik P, Bu X, Kelly TN, Mills KT, He H, Chen J, Whelton PK, and He J. Systolic blood pressure reduction and risk of cardiovascular disease and mortality: a systematic review and network meta-analysis. JAMA Cardiol 2017 Jul 01;2(7):775-781

[131] Fogg, B. Fogg behavior model. URL: https://behaviormodel.org.