



Efficient and Direct Inference of Heart Rate Variability using Both Signal Processing and Machine Learning

Yuntong Zhang
University of Texas at San Antonio
USA

Jingye Xu
University of Texas at San Antonio
USA

Mimi Xie
University of Texas at San Antonio
USA

Dakai Zhu
University of Texas at San Antonio
USA

Houbing Song
University Of Maryland
USA

Wei Wang
University of Texas at San Antonio
USA

Abstract

Heart Rate Variability (HRV) measures the variation of the time between consecutive heartbeats and is a major indicator of physical and mental health. Recent research has demonstrated that photoplethysmography (PPG) sensors can be used to infer HRV. However, many prior studies had high errors because they only employed signal processing or machine learning (ML), or because they indirectly inferred HRV, or because there lacks large training datasets. Many prior studies may also require large ML models. The low accuracy and large model sizes limit their applications to small embedded devices and potential future use in healthcare.

To address the above issues, we first collected a large dataset of PPG signals and HRV ground truth. With this dataset, we developed HRV models that combine signal processing and ML to directly infer HRV. Evaluation results show that our method had errors between 3.5% to 25.7% and outperformed signal-processing-only and ML-only methods. We also explored different ML models, which showed that Decision Trees and Multi-level Perceptrons have 13.0% and 9.1% errors on average with models at most hundreds of KB and inference time less than 1ms. Hence, they are more suitable for small embedded devices and potentially enable the future use of PPG-based HRV monitoring in healthcare.

Keywords: Heart Rate Variability, Machine Learning, Photoplethysmography, Signal Processing

1 Introduction

Heart Rate Variability (HRV) measures the variation of the time intervals of consecutive heartbeats and is a major indicator for health conditions, such as coronary artery disease, heart failure, hyperlipidemia, and hypertension [40]. HRV

is traditionally measured using electrocardiographic (ECG) devices, which record the heart's rhythm. However, ECGs can be expensive and require attaching several electrodes to human bodies, which can be inconvenient to use.

As an alternative to ECG, Photoplethysmography (PPG) sensors, which monitor light signal changes in blood flows, can also be utilized to measure heart rate (HR) and HRV [37, 45]. A PPG sensor can be placed on the human skin to provide HR/HRV readings and is more convenient to use. Many studies demonstrated the potential of PPG sensors in HRV monitoring [14, 39]. However, these studies usually face the following four limitations, restraining their application to small embedded devices and healthcare in the future.

First, some of these studies relied on only signal processing for HRV estimation [6, 45], which may negatively affect the inference accuracy. PPG sensor signals typically suffer from large noises, particularly, the noises from motion artifacts (MA). To infer HRV with good accuracy, these noises must be removed or reduced. Although signal processing techniques can remove many noises (including MA), these relatively static techniques may not be able to handle all types of noises, leading to low HRV accuracy.

Second, many studies also only employed machine learning (ML) techniques to infer HRV [11, 14]. Although many ML models by themselves are sophisticated enough to handle all types of noises, the resulting models can be too large and/or too slow for small embedded devices. Moreover, training and tuning ML models with raw PPG data can be quite challenging. ML models without enough tuning may also have low accuracy.

Third, some prior studies also focused on the inference of the interval lengths between consecutive heart beats [14, 41], which are known as RR intervals [13, 18]. That is, these studies do not direct infer HRV metrics, such as SDNN or RMSSD (more in Section 2). This indirect inference strictly follows HRV's definition. However, due to error amplification, this indirect inference can significantly increase the errors of the final HRV estimations.

Fourth, many prior studies also relied on small datasets. For example, a popular PPG dataset, the IEEE Signal Processing Cup (ISPC) dataset, has PPG signals recording lasted for only five minutes [45]. However, a typical HRV inference requires



This work is licensed under a Creative Commons Attribution International 4.0 License.

CHASE '23, June 21–23, 2023, Orlando, FL, USA

© 2023 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-0102-3/23/06.

<https://doi.org/10.1145/3580252.3586971>

an ECG monitoring length of half to five minutes [1]. Hence, it is difficult to conduct HRV inference studies with short recordings, especially for studies with neural networks.

We have studied ML-based HR estimation using PPG in our prior work [43]. In this paper, our goal is to design an HRV inference methodology that can provide high accuracy for resource-constrained embedded devices. To achieve this goal, we designed a *compound and direct* method that combines signal processing and ML to directly infer HRV (i.e., RMSSD/SDNN). More specifically, we first employed signal processing to remove outliers and noises from raw PPG signals and convert the PPG signals into rough HR readings and HRV readings. These rough HR/HRV readings are then fed into an ML model to infer RMSSD/SDNN. Applying ML after signal processing provides additional/better noise removal, and hence, more accurate HRV estimations. Applying signal processing before ML avoids the need for large and slow ML models. Moreover, the direct inference of RMSSD/SDNN, instead of inferring RR intervals as a proxy, further improves accuracy.

To explore the impact of various ML algorithms, we also evaluated different ML models, including Decision Tree (DT), Random Forest (RF), K-nearest neighbor (KNN), Support vector machines (SVM), and Multi-layer perceptron (MLP). To provide more reliable results, we also collected a new dataset of PPG signals, with ECG readings as ground truth. This dataset contains three 2-hour-long PPG/ECG traces for one human subject performing different activities, including office work, sleeping, and sitting.

Evaluation results show that our compound and direct method has 3.5% to 25.7% errors for various activities and monitoring lengths. Both the lowest 3.5% error for RMSSD and the lowest 5.1% error for SDNN were obtained with a monitoring length of 300 seconds per HRV estimation. Note that, 300-second HRV monitoring can be used for caring for chronic renal failure and diabetes [1], showing the healthcare potential of PPG-based HRV monitoring. The evaluation results also show that our method is significantly more accurate than the signal-processing-only and ML-only methods.

Moreover, the model exploration showed that DT and MLP models are usually smaller with good accuracy, making them more suitable for small embedded devices. DT models have an average error of 13.0% and are usually less than 10KB with inference time less than 10 μ s. MLP models have an average error of 9.1% and are less than 469KB with inference time less than 1ms. These results corroborate the "Rashomon" theory [32] that, for some problems, there exist simple models with good accuracy and meet special requirements, such as the limited memory size of embedded devices in this case.

The contributions of this paper include,

- The compound and direct HRV inference methodology combines signal processing and ML to directly infer RMSSD/SDNN to achieve high accuracy with small and fast ML models.

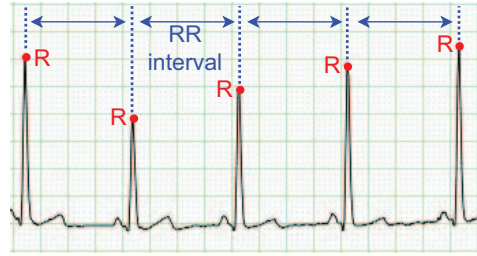


Figure 1. An example of ECG waveform showcasing R peaks (red dots) and RR intervals.

- A systematic exploration of different ML algorithms to study their impact on the accuracy, model size, and time on HRV inference. This exploration showed that Decision Trees and MLP can achieve high accuracy with small/fast models suitable for embedded devices.
- A comprehensive PPG/ECG dataset to study HR/HRV inference, which contains traces of different activity intensities lasting for 2 hours.

The rest of the paper is structured as follows. Section 2 discusses the background on HRV and the motivation of our work. Section 3 presents our compound and direct method. Section 4 presents evaluation results. Section 5 discusses related work, and section 6 concludes the paper.

2 Background and Motivation

2.1 HRV from ECG

Human heart rate, when measured as beat-to-beat intervals, is not constant and varies over time [31]. This variation, commonly known as Heart Rate Variability (HRV), is an effective indicator of various health and mental problems [40]. The traditional medical device to measure HR and HRV is ECG. ECG records heart activity utilizing electrodes placed at certain skin spots on the human body and produces an electrocardiogram, which is a graph that shows the heart's activity over time. Electrocardiogram contains the QRS complexes information, which is an important waveform in an electrocardiogram that shows the spread of a stimulus through the ventricles [13, 18]. R peaks, which roughly represent heartbeats, can be computed from the QRS complex. The intervals between R peaks are called RR intervals. Figure 1 gives an example output from an ECG device that shows the heartbeats and RR intervals. Note that, normal RR intervals are also called NN intervals in the literature [40].

HRV is defined as the variation of the RR intervals within a time period. HRV is an important health indicator because it represents the adaptive ability of the heart to unpredictable changing circumstances. There is no one standard or best method to calculate HRV [21]. In this work, we focus on two commonly used time-domain linear measures for HRV - the Standard Deviation of RR intervals (SDNN) and the

Root Mean Square of Successive Differences (RMSSD) [20]. SDNN is usually recommended for overall HRV estimation and represents both sympathetic and parasympathetic modulation of heart rate, whereas RMSSD is recommended for estimating short-term components of HRV and represents parasympathetic activity [24, 40].

An SDNN or RMSSD is calculated from the RR intervals in a chosen time window, which is usually between 0.5 and 5 minutes [1], and a sequence of HRVs over 5 minutes to 24 hours are typically used in medical practice [24]. The definitions of SDNN and RMSSD are,

$$\text{SDNN} = \sqrt{\frac{\sum_{i=1}^N (RR_i - \overline{RR})^2}{N}}, \quad (1)$$

$$\text{RMSSD} = \sqrt{\frac{\sum_{i=1}^{N-1} (RR_{i+1} - RR_i)^2}{N-1}}, \quad (2)$$

where RR_i is the i^{th} RR interval, \overline{RR} is the average, and N is the number of RR intervals within the chosen time window.

Although the ECG produces accurate HRVs, attaching electrodes to the human body can be inconvenient for long-term monitoring [28, 35]. Therefore, in this work, we focus on using PPG instead of ECG. Nonetheless, we did use ECG to collect reliable HRV readings as the groundtruth to train and evaluate our PPG-based HRV solutions.

2.2 HRV from PPG

PPG sensors are popular due to their non-invasive nature. They are usually attached to human skins at certain locations, such as fingertips, earlobe, and wrist [9]. These sensors utilize infrared light that penetrates the skin to detect changes in the blood circulation – if there is a change in the blood flow, the intensity of the infrared light also changes [30]. Hence, by monitoring the light changes, PPG sensors can detect blood flow changes, which in turn, can be used to infer HR/HRV.

The main challenge to using PPG for HR/HRV monitoring is light signal noises, especially the motion artifact (MA), which represents the signal noises due to body/hand movements [15]. There are also noises from the environment [9, 15] or from the inherent sensor inaccuracy/bias (e.g., sensor sensitivity and calibration issues). Accurate HR/HRV monitoring requires the removal or reduction of these noises [9, 45].

2.3 Motivation

Although there have been many studies on applying PPG in HRV inference, these studies were usually limited by their methodology and/or by small datasets. To illustrate these limitations, and as a motivation for this study, we explored the HRV inference accuracy of a signal-processing-only method [45] and ML-only method with Convolutional Neural network (CNN) based encoder-decoder [14] using the popular ISPC dataset.

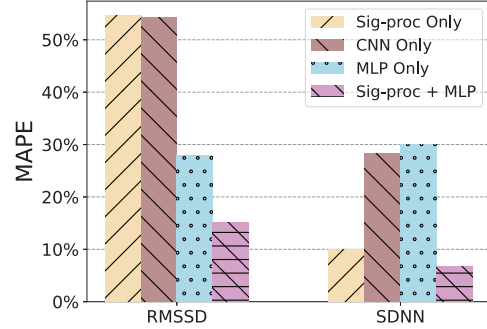


Figure 2. HRV estimation errors of the signal-processing-only [45], CNN-only [14], MLP-only, and our compound and direct ("Sig-proc+MLP") methods, using the ISPC dataset.

RR MAPE	RMSSD MAPE	SDNN MAPE
1%	10.24%	4.99%
2%	27.83%	14.14%
3%	41.55%	23.87%
4%	52.58%	33.19%
5%	61.18%	41.97%

Table 1. RR estimation error amplification after converting to RMSSD and SDNN.

HRV	CNN Only	MLP Only	Sig-proc + MLP
SDNN	195KB	47.9MB	47.9KB
RMSSD	195KB	122.5MB	49.8KB

Table 2. CNN or MLP model sizes of the CNN-only [14], MLP-only, and our compound and direct ("Sig-proc+MLP") methods, using the ISPC dataset.

Figure 2 gives the MAPEs (mean absolute percentage error) of these two methods. For the signal-processing-only method, we reproduced its processing procedures to generate estimations of HRV, which were then compared with the ECG ground truths in the ISPC dataset to calculate the MAPEs. For the CNN-only method, its MAPEs were directly calculated based on results reported in its paper. As Figure 2 shows, both methods have high errors for HRV estimations. Especially for RMSSD, both methods have over 50% error.

We observe four reasons that cause these high errors.

1. First, the static signal processing cannot always effectively remove all noises in the PPG signals. Therefore, for RMSSD, which measures the short-time components of HRV, the remaining signal noises could significantly degrade the accuracy of the signal-processing-only method. Interestingly, the signal-processing-only method's SDNN estimation accuracy is less affected,

as SDNN represents long-term variability and is less sensitive to the remaining signal noises.

2. Second, the CNN-only method suffers from error amplification. This method does not directly estimate RMSSD or SDNN. Instead, it estimates RR intervals, which are then converted to RMSSD/SDNN using Equations (1) and (2). However, this conversion amplifies the estimation error. Table 1 illustrates this error amplification, where we generated five sets of RR estimations with random errors based on certain average errors (MAPE), converted them into HRV (RMSSD/SDNN), and evaluated the errors of the converted HRV. Table 1 shows that a small 3% MAPE in RR estimations amplifies to 41.55%/23.87% error for RMSSD/SDNN.
3. Third, for the CNN-only method, although it can remove most of the noises in theory, the small ISPC dataset does not provide enough data for the CNN model to learn the noise removal completely. The ISPC dataset only has PPG signals from five minutes of monitoring, whereas a single HRV reading requires half to five minutes [1]. This small dataset further contributes to the CNN model's high errors for both RMSSD and SDNN estimations in Figure 2.
4. Fourth, the model used in the CNN-only method could be too small and not sophisticated enough to process noisy signals. Table 2 shows that this CNN model is only 195KB (42657 trainable parameters). As shown later, inferring HRV with raw noisy PPG signal would require complex neural network models of tens or hundreds of MBs, as the noise removal computation is typically nonlinear and non-polynomial. Large models, however, are unsuitable for small wearable devices with only hundreds of KBs of on-chip memory.

Based on the above observation, our *hypothesis* is that a *compound and direct method* that combines signal processing and ML, and directly estimates RMSSD/SDNN, can achieve both high accuracy and small ML model size. To verify this hypothesis, we trained MLP-based models to directly estimate RMSSD and SDNN using the signal-processed ISPC's PPG data. The accuracy and model size of our method are also given in Figure 2 and Table 2 under the label "Sig-proc+MLP". As Figure 2 shows, our method had lower error than both the signal-processing-only and ML-only methods. Table 2 also shows that this compound and direct method had small model sizes of less than 50KB because many signal noises are already treated by signal processing.

For the sake of comparison completeness, we also trained MLP models using the original PPG signals from the ISPC dataset to directly estimate RMSSD/SDNN. That is, we also compared an ML-only method with MLP models. These MLP models went through hyperparameter tuning, and the errors of the most-accurate MLP models are reported in Figure 2 under the label "MLP-only", which shows that this MLP-only

method has lower accuracy than our compound and direct method, mainly due to the small dataset. Moreover, as Table 2 shows, the sizes of the MLP-only models are much larger than our compound models, due to the need of relying on pure neural networks to remove all noises.

In summary, the above results show that our hypothesis is likely to be valid, although more data are required to further validate this hypothesis. In the rest of this paper, we will present the details of our methodology for data collection and HRV estimation.

3 Compound and Direct HRV Estimation

In this section, we present our signal processing and machine learning combined method for direct HRV estimation.

3.1 Data Collection

A single HRV estimation typically requires 30 seconds to 5 minutes of PPG/ECG monitoring [1]. Moreover, typically, a sequence of five minutes to 24 hours of HRV data is used in medical practice [40]. Therefore, studying HRV estimation requires hours of PPG/ECG monitoring data to provide enough data points. As there lack of such public datasets, our first task in this research was to collect longer traces of PPG and ECG data.

One subject participated in this data collection. The subject had a PPG sensor attached to the fingertip and an ECG monitor attached to the chests at the same time. The PPG data were collected as features/inputs for the HRV inference, whereas the ECG data were used as the labels for model training and as the groundtruth in model testing/evaluation. To reduce the energy consumption caused by the PPG sensors, we collected PPG readings at a frequency of 25Hz, rather than the 125Hz used by the ISPC dataset and other studies [4, 10, 27, 44].

The subject conducted three activities during the data collection, including sitting, sleeping, and office work. The office work activity includes actions such as working in front of a computer, walking, and drinking water. For each activity, more than two hours of PPG/ECG data were collected, which gave about 180193 to 180271 PPG readings per activity.

3.2 Workflow for HRV Inference

Figure 3 gives the overall workflow of our HRV inference methodology. This workflow includes three major steps, PPG data collection, signal processing, and machine learning-based HRV inference. The following paragraphs provide a detailed description of each step.

3.2.1 Step 1: PPG Monitoring. As our HRV inference is based on PPG sensors, the first step is to collect the PPG light signals reflected from the blood flow. Here, we employed a sampling rate of 25Hz, i.e., 25 signals are collected every second. This sampling rate is lower than the 125Hz used by many prior studies [4, 10, 27, 44], and it is employed to reduce the power consumption of sensing, as prior work [6]

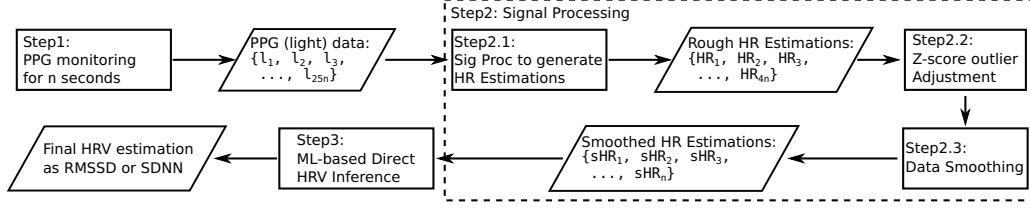


Figure 3. The workflow of our compound and direct HRV inference.

has shown that high sampling frequency incurs high power usage. Low sampling frequency, however, may negatively affect HRV inference accuracy. As discussed later, we rely on signal processing and machine learning to compensate for this negative accuracy impact. Let $\{l_1, l_2, l_3, \dots, l_{25n}\}$ denotes the sequence of the $25n$ light signals over n seconds. These signals are fed to step 2 for processing.

3.2.2 Step 2: Signal Processing. The primary goal of our signal processing is to remove the noises due to motion artifacts with outlier adjustment and data smoothing. The secondary goal of our signal processing is to generate HR estimations to be used later as features for ML models.

Step 2.1: Signal Processing to Generate HR Estimations. In this step, we convert the PPG light signals collected in Step 1 into four heart rate estimations per second using signal processing. This processing essentially applied a peak detection algorithm to the PPG signals. These peaks can be viewed as "heartbeats", and hence, their counts can be used to estimate HR [2]. After this conversion, there are $4n$ HR estimations over n seconds, denoted as $\{HR_1, HR_2, HR_3, \dots, HR_{4n}\}$.

This conversion serves two purposes. First, this conversion simplifies the noise removal (i.e., outlier adjustment and smoothing) conducted later. Motion artifacts typically affect a sequence of PPG light signals, and it can be challenging to distinguish erratic signals from real HR fluctuations when working on raw signals. Converting to HR estimations reduces the number of data points, making it easier for noise removal. Second, these converted HRs are also used as the features of our ML models in Step 3.

Step 2.2: Z-score Based Outlier Adjustment. Many noisy signals can be simply viewed as outliers. Therefore, outlier identification algorithms can be used to remove these signal noises. More specially, we applied the popular Z-score-based outlier filter algorithm [23, 33].

The Z-score filter identifies the outliers by picking out the data points that deviate the most from the mean. Concretely, consider a large sequence of HR estimations, $\{HR_1, HR_2, \dots\}$, with mean μ and standard deviation δ . If the difference between a data point HR_i and μ (i.e., $|HR_i - \mu|$) is larger than a threshold, then HR_i can be viewed as an outlier. This threshold is typically defined based on the standard deviation δ . Particularly, the threshold is defined as $z_score \times \delta$. In this work, we used a z_score of 3, following common practices [36, 42].

That is, if the difference between HR estimation HR_i and the mean μ is larger than 3δ , then HR_i is deemed as an outlier.

We do not remove outliers because HR estimations are time series, and removing estimations would create "holes" within the time series. Consequently, after the outliers are identified, their values are just adjusted to be the average of their neighboring estimations. That is, given an outlier HR_i , its value is adjusted to be $\frac{HR_{i-1} + HR_{i+1}}{2}$.

Step 2.3: Data Smoothing. The above outlier adjustment only identifies data points with large noises, i.e., data points deviate greatly from the overall mean. However, there could still be HR estimations that deviate significantly from local HR averages. These deviating HR estimations usually manifest themselves as abnormal local peaks/valleys. These local peaks/valleys are usually caused by signal noises, because normally, a person's heart rate does not increase (or drop) abruptly and drops (or increases) back within one second.

To remove these local noises, we apply moving average data smoothing. More specifically, this data smoothing converts the four rough HRs within a second into one HR estimation per second by computing their averages to reduce the impact of the abnormal local peaks and valleys. After the smoothing, there are n smoothed HR estimations over n seconds, denoted by $\{sHR_1, sHR_2, sHR_3, \dots, sHR_n\}$.

3.2.3 Step 3: ML-based HRV Inference. The last step of our methodology employs ML to infer HRV. As stated previously, the ML models are trained to serve two purposes simultaneously. First, they are trained to further remove/reduce the noises that cannot be filtered by signal processing. These additional noises may include long motion artifacts that affect several seconds of PPG signals, the sensor bias (e.g., sensor sensitivity and calibration issues), and the errors due to our low sampling frequency. Second, with further reduced noises, these ML models estimates the final HRV. These models directly infer SDNN/RMSSD, instead of the RR intervals.

The main input features to the ML models are the smoothed HR estimations over n seconds from Step 2, i.e., the vector $\{sHR_1, sHR_2, sHR_3, \dots, sHR_n\}$. These smoothed HR estimations are also used to derive a rough HRV (SDNN or RMSSD) estimation, denoted by $rHRV$, using the Equations (1) and (2). This rough HRV is used as a constructed feature for our ML models. As shown with Equations (1) and (2), computing HRV requires exponentiation and square root operations,

which may require large ML models to simulate. Therefore, employing a computed rough HRV estimation as a feature can potentially reduce the trained model size.

In summary, the features of our ML models are the vector, $\{sHR_1, sHR_2, sHR_3, \dots, sHR_n, rHRV\}$. The output of our models is the HRV estimation for the past n seconds in SDNN or RMSSD. Note that, n represents the number of seconds of PPG monitoring, which is a tuneable parameter depending on the HRV use case. In the experimental evaluation (Section 4), we evaluated different values for n .

3.3 Model Training

During our PPG data collection (described in Section 3.1), the subject also had an ECG attached to collect groundtruth HRV readings. The groundtruth HRVs are used both as the labels in the training data and validation/testing data.

We partitioned the collected data into training and testing datasets with a split of 80% and 20%. All models were also optimized with hyperparameter tuning [5] to find the model with the best accuracy. The specific hyperparameters for each type of model are given in Section 4.1.2. Note that, we limited hyperparameter search space to avoid generating models that are too large for small embedded devices.

4 Experimental Evaluation

This section presents the evaluation results, focusing on the HRV inference accuracy, ML model sizes, and inference time.

4.1 Experiment Setup

4.1.1 Hardware used in Data Collection and Inference.

Our dataset contains one channel of PPG data. The subject simultaneously wore the PPG sensor on the fingertip and ECG electrodes. The PPG sensor is connected to a Raspberry Pi 4B through an i2C bus to record the data. The hardware components are: 1) a Raspberry Pi 4B with 4 GB RAM; 2) a Maxim MAXREFDES117 HR monitor with a MAX30102 PPG sensor; and 3) a TLC5007 Dynamic ECG. PPG data and ECG data are synchronized according to their timestamps. The Raspberry Pi is also used to measure the inference latency of the signal processing and ML models.

4.1.2 ML Models and Hyperparameter Configurations.

DT, RF, KNN, and SVM models were implemented using Scikit-learn [26] and MLP was implemented using Keras. ML models were saved with *joblib*. We used the random search function *RandomizedSearchCV* in Scikit-learn and *RandomSearch* in Keras-tuner to tune the model hyperparameters. The hyperparameters are: 1) For DT, the maximum depth of the tree ranges from 3 to 20; 2) For RF, the number of trees is between 2 to 128, and the maximum tree depth is between 3 to 20; 3) For KNN, the number of neighbors is between 2 to 30, and the distance can be Manhattan or Euclidean; 4) For SVM, the kernel may be among RBF, sigmoid and polynomial, and regularization (i.e., C) may be between

0.00001 to 10; 5) For MLP, there are 1 to 5 hidden layers, each with 1 to 100 neurons, and the activation function can be *relu* or *tanh*.

4.1.3 Metrics. For HRV estimation accuracy evaluation, we used the metric MAPE (Mean Absolute Percentage Error) between the HRV estimations and groundtruths. Given m HRV estimations, the definition of MAPE is,

$$MAPE = \frac{100\%}{m} \sum_{i=1}^m \left| \frac{HRV_{est,i} - HRV_{true,i}}{HRV_{true,i}} \right|. \quad (3)$$

For model size evaluation, we report the sizes in KB (kilobytes) or MB (megabytes). For inference time evaluation, we report the average time it takes to make an inference in microseconds (μs) or milliseconds (ms).

4.2 Accuracy Evaluation

Figure 4 and Figure 5 give the MAPE of our compound and direct methods, using different types of ML algorithms at different monitoring lengths (i.e., the n seconds in Section 3.2.3). The figures also give the errors of the signal-processing-only method ("Sig-proc Only"). The following paragraphs discuss these accuracy results in detail.

4.2.1 Overall Accuracy of Our Method.

Figure 4 and Figure 5 show that our compound and direct prediction methods usually had errors between 3.5% to 25.7%. The highest error was 25.7%, which was for the KNN model for SDNN estimation under the office work scenario with 30 seconds monitoring length (Figure 5a). The lowest error was only 3.5%, for the MLP model for RMSSD inference under sit scenario with 300 seconds monitoring length (Figure 4c).

For the majority of the models, the errors of our method were less than 20%. Figure 6 also shows that the overall average MAPEs for different ML models (overall activities for both RMSSD and SDNN) are all less than 13.2%. MLP models have the lowest overall average MAPEs of only 9.1%. These results show that our compound and direct method has high accuracy for HRV estimation.

4.2.2 Comparison with Signal Processing Only.

Figure 4 and Figure 5 also show that our method was usually more accurate than the signal-processing-only method ("Sig-proc Only"). In the case of RMSSD estimation (Figure 4), the signal-processing-only method usually had errors above 20%, whereas our method's errors are usually less than 20%.

In the SDNN estimations (Figure 5), the signal-processing-only method had lower errors than their RMSSD estimations, because SDNN measures long-term HRV and is less sensitive to signal noises. We had a similar observation for the ISPC dataset as discussed in Section 2.3. Nonetheless, our method usually still had lower errors than the signal-processing-only method for SDNN estimations.

The only exceptions were for SDNN estimations in sit scenario with 240/300 seconds monitoring lengths (Figure 5c, where our method with some ML models (e.g., SVM) had

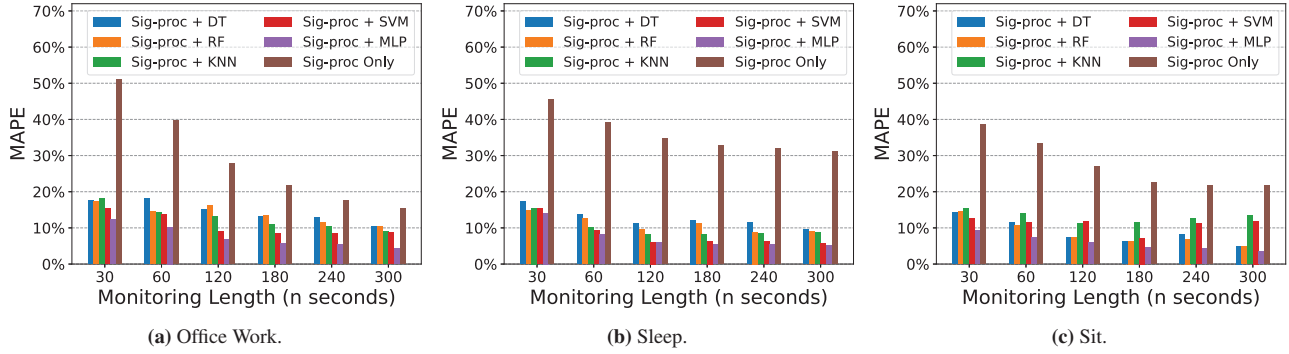


Figure 4. MAPEs for HRV/RMSSD estimations for different activities.

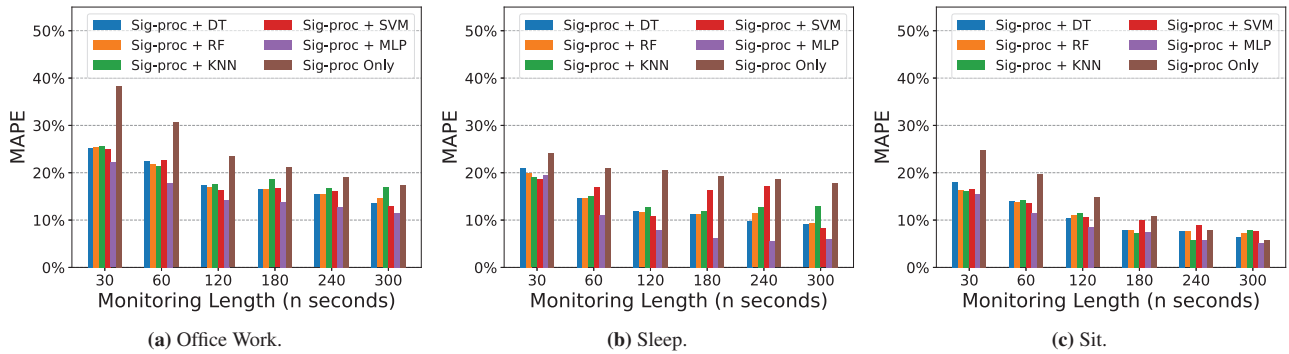


Figure 5. MAPEs for HRV/SDNN estimations for different activities.

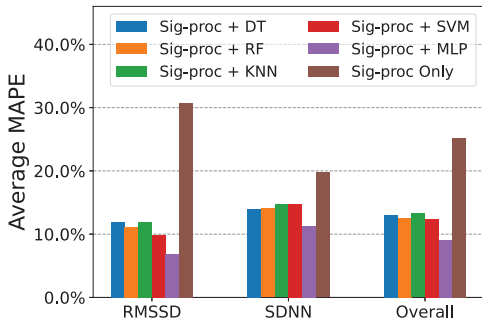


Figure 6. Average MAPEs across all activities and all monitoring lengths.

higher errors than the signal-processing-only method. However, the error difference was small – only 2.2% at most. Moreover, our method with the MLP model was still more accurate than signal-processing-only in these cases.

Overall, Figure 6 shows that the signal-processing-only method has an overall average MAPE of 25%, which is higher

than the overall average MAPEs (about 12%) of our compound and direct method using any ML model.

4.2.3 Traces of HRV Estimation. Figure 7 and Figure 8 give the traces of the RMSSD HRVs from the groundtruth (ECG), signal-processing-only method, and our method with MLP model for the sleep activity data. These traces illustrate that our MLP models reduce at least two types of noises after signal processing. The first type of noise is mostly errors due to either sensor bias or low sampling frequency. For example, for the 5600th to 5750th HRVs in Figure 7, signal-processing HRVs had similar fluctuation trends as the ECG HRVs, but they deviate by roughly 10. The MLP model, however, was trained to correct this "deviation" and produced more accurate HRVs. This same issue can also be observed for the SDNN estimations in Figure 8. The second type of noise is usually from motion artifact noises that affect a longer period. For example, for the 5900th to 6200th HRVs in Figure 7, signal-processing HRVs were flat then dropped, whereas the ECG HRVs were sharply increasing. These longer errors were also detected and corrected by our MLP model to produce HRVs increasing from 38 to 50, similar to the ECG.

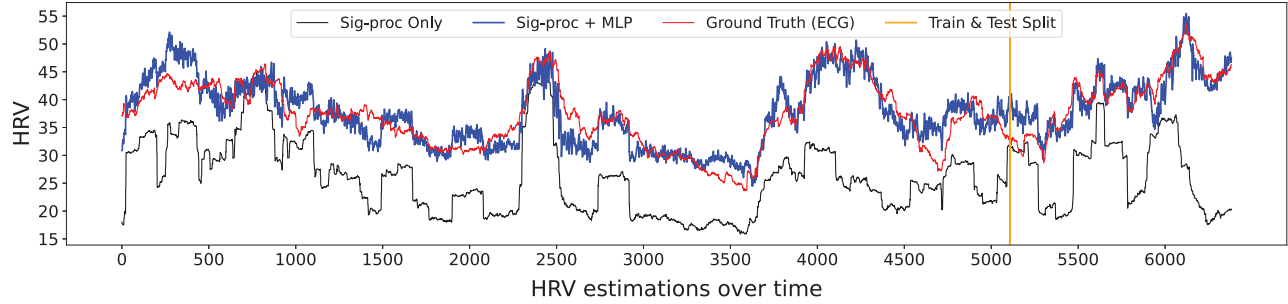


Figure 7. RMSSD estimation trace (sleep activity, monitoring length of 180sec). The yellow line separates training and test data.

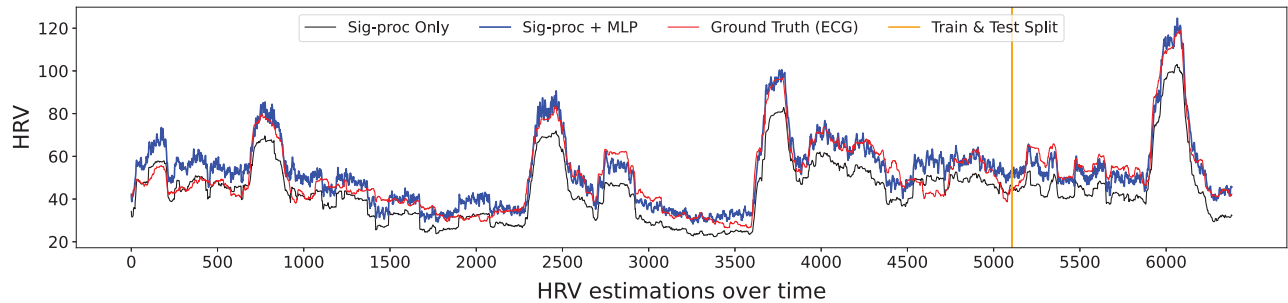


Figure 8. SDNN estimation trace (sleep activity, monitoring length of 180sec). The yellow line separates training and test data.

The same conclusion can be drawn from the traces for other models, activities, and monitoring lengths. However, due to space limitations, these traces are omitted.

4.2.4 Impact of Type of Activity and HRV Metric. Across the results for the three activities in Figure 4 and Figure 5, HRV estimations for sit had the lowest MAPEs. These low errors were because when the subjects sat, they had little movement, hence, lower motion artifacts. In this work, we built one ML model for each activity. However, if only one ML model is built to cover all activities, then these MAPE differences suggest that more sensors/features (e.g., accelerometer or gyroscope) may be needed to conduct noise correction differently for different activities.

Moreover, Figure 6 also show that our method’s average MAPEs only differ slightly for RMSSD and SDNN estimations, indicating that our method works for both short-term or long-term HRV monitoring. However, the signal-processing-only method had considerably lower errors for SDNN (long-term) estimations than RMSSD (short-term) estimations.

4.2.5 Accuracy Impact of Model Types. Figure 6 shows that MLP models were generally more accurate than other ML models in our method, with average MAPEs for RMSSD, SDNN, and "overall" being only 6.9%, 11.3%, and 9.1%, respectively. Nonetheless, the differences in average MAPEs among model types in Figure 6 are less than 4%. These similar MAPEs match the recently proposed "Rashomon" theory [32],

which states that there could be multiple ML models having similar accuracy for the same dataset. A group of models with similar accuracy implies that it is beneficial to conduct ML model exploration to search for models that fit certain non-functional requirements, such as low model sizes and fast inference time in the embedded applications.

4.2.6 Accuracy Impact of Monitoring Length. Figure 4 and Figure 5 also show that another factor that impacts the accuracy of our method is the monitoring length n . That is, the longer the monitoring length, the lower the HRV estimation error. And the HRV estimations for the monitoring length at 300 seconds usually had the lowest errors. An HRV estimation for longer monitoring is usually less susceptible to a few noisy PPG signals, and therefore, it tends to have lower errors.

Note that, the required HRV monitoring length depends on the use case [40]. For example, 300-second monitoring is applied for caring for chronic renal failure and diabetes [1]. The fact that errors vary with monitoring lengths also suggests the applicability of PPG-based HRV monitoring to medical use needs to be evaluated case by case.

4.2.7 Comparison with the ML-only Method. We also experimented with the ML-only method to directly infer HRV using the original PPG data collected from the sensors. However, we were not able to obtain any ML-only HRV models with good accuracy, or even meaningful estimations. These models typically have accuracy similar to, sometimes even

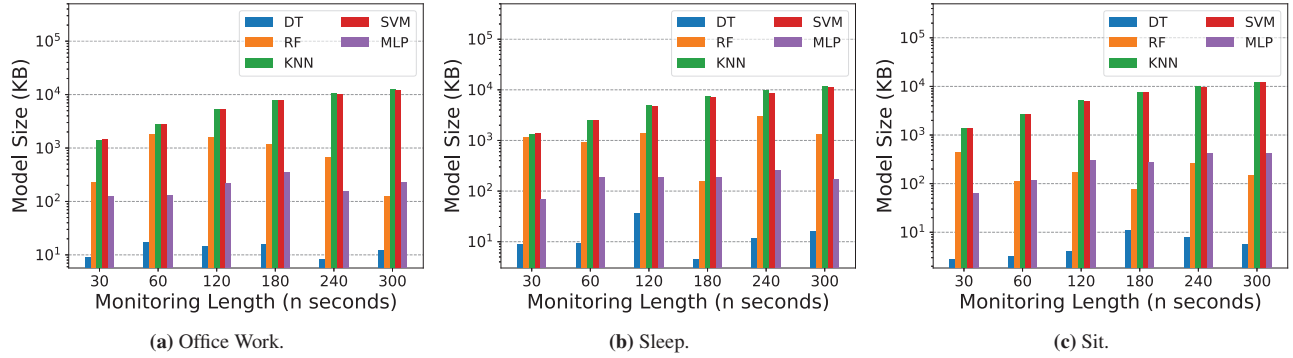


Figure 9. Sizes of HRV/RMSSD models for different activities.

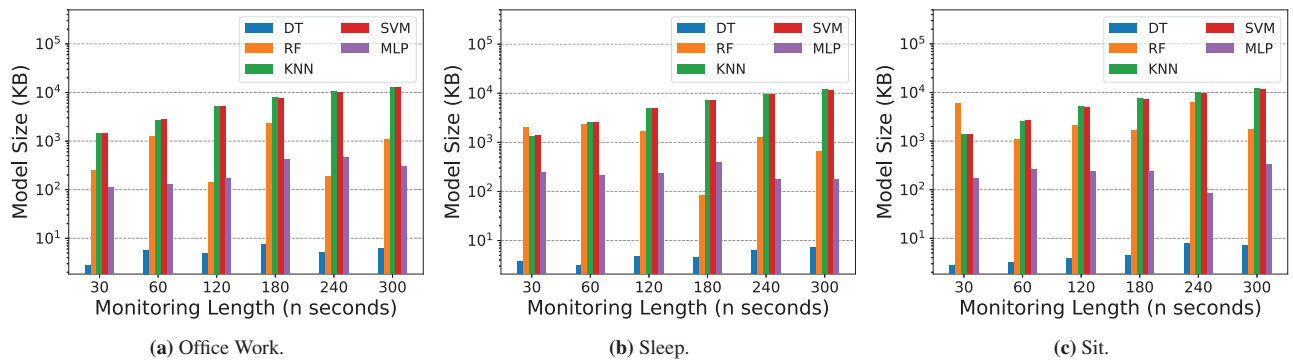


Figure 10. Sizes of HRV/SDNN models for different activities.

worse than, the signal-processing-only method. For the case of MLP, the HRV estimations produced by each MLP model are mostly the same value, making the estimations practically useless. These MLP models also have average sizes of 6.7MB and maximum size of 26.6MB, larger than the MLP models in our compound method.

We believe the main cause of the low accuracy was the small hyperparameter search space. Recall that we limited the hyperparameter search space to limit the ML model sizes, which also limits the model complexity. However, it typically needs very complex ML models to infer HRV using only the original PPG signals. Because of the poor results of the ML-only HRV models, we did not include them in the paper.

4.3 Model Size Evaluation

Figure 9 and Figure 10 give the sizes of the ML models generated by our direct and compound HRV inference method. As both figures show, the ML model sizes ranged from 2.8KB to 12.4MB. The figures also show that the model sizes are less affected by the type of activities and the HRV metric (RMSSD or SDNN). Monitoring lengths have a higher impact on the

model size – the longer the monitoring length, the more input features, and hence, larger models.

The largest factor for model size is the type of model. RF, KNN, and SVM models are usually large and approaching 10MB. As the input HRs to the ML models are generated from PPG signals, they are usually "messy" and lack a clear "pattern." Hence, RF/KNN/SVM requires large numbers of internal parameters to learn their "patterns." The MLP models, however, are smaller – most MLP models are below 200KB, with the smallest model being 63.4KB (Figure 9c at 30s) and the largest model being 468.5KB (Figure 10a at 240s). DT models are even smaller, – most DT models are below 10KB, with the smallest model being 2.8KB (Figure 9c at the 30s) and the largest model is 35.4KB (Figure 9b at the 120s). Considering DT and MLP models can fit in on-chip memory of hundreds of KB, they are better candidates for deploying to tiny embedded devices.

4.4 Inference Time Evaluation

Figure 11 and Figure 12 give the inference time of our compound and direct method. As both figures show, the inference time for all models was less than 7.3ms (max 7295 μ s). DT

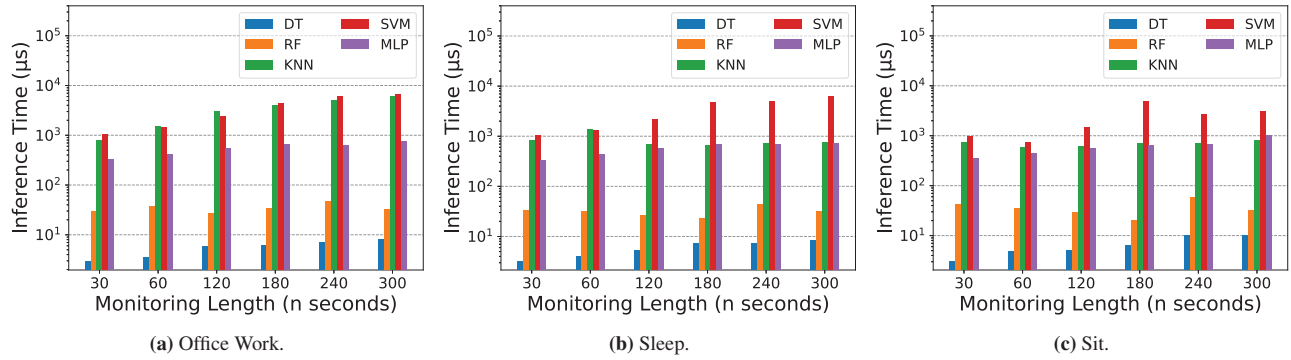


Figure 11. ML model inference time for HRV/RMSSD estimations for different activities.

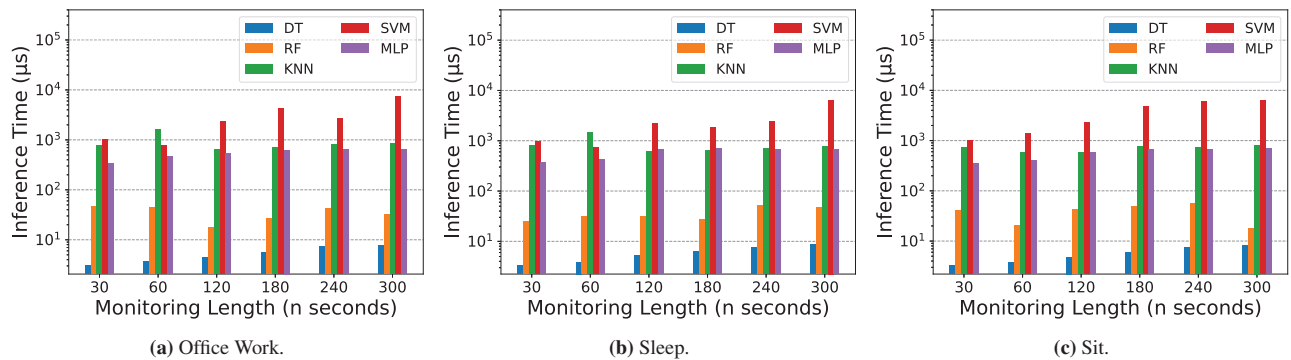


Figure 12. ML model inference time for HRV/SDNN estimations for different activities.

models are the fastest models due to their small sizes – the inference time for DT models is between only $2.9\mu s$ to $10\mu s$. The inference time of RF models is also fast, ranging between $17.6\mu s$ to $57.2\mu s$. The SVM models are the slowest due to their large sizes, and their inference time ranges from $0.7ms$ to $7.3ms$. Nonetheless, even $7.3ms$ is fast enough for HRV inference, indicating that all models under our methodology are fast enough for HRV monitoring with embedded devices.

The signal processing takes $65.74ms$ on average and $68.31ms$ at maximum, which is significantly slower than ML model inference.¹ Nonetheless, the signal processing is still fast enough for HRV monitoring. Note that, the total computation time for our method includes both the signal processing time and ML inference time.

5 Related Work

Although there are many existing works on HR monitoring [4, 7, 10, 25, 43, 45], only a few studied HRV monitoring.

¹This $65.74ms$ is also the average processing time for the signal only method.

5.1 PPG HRV Monitoring with Signal Processing

Ghamari et al. proposed a signal processing algorithm including High-Pass and low-Pass Filters to detect R peaks [17]. Srinivas et al. proposed a signal processing method including moving average and FFT to measure HRV based on the PPG wave [34]. However, both of them did not evaluate the accuracy of HRV estimations. Wang et al. proposed an algorithm to estimate RR intervals from a smartwatch PPG sensor and accelerometer [38]. They calculated HRV measurements, such as SDNN and RMSSD. However, no quantitative accuracy was reported. Blake et al. devised an HRV estimation hardware that contains a PPG sensor, an accelerometer, a Bluetooth module, and a battery [8]. This study only compared the HRV readings from a chest strap instead of ECG with no quantitative report on accuracy. Janković and Stojanović designed a signal processing algorithm including a low-pass filter, Sum Slope Function, and a peak extraction function to find R peaks [19]. They collected data for 8 minutes in the experiment and obtained one HRV for each trace, and compared them with ECG HRV. However, it was unclear which HRV metric they used. Bhowmik et al. proposed an algorithm including wavelet denoising, trend removal, and

peak extraction to detect R peaks in PPG signals [6]. They found that a 100Hz PPG sampling rate is not suitable for a smartwatch due to high power consumption and chose 25Hz. Inspired by this work, we also choose 25Hz as the PPG sampling rate to save energy. Saadeh et al. employed earlobe-attached PPG and wavelet decomposition and moving average filters to estimate RR intervals [29].

Note that, due to the nature of PPG light signals, the above prior studies all estimated RR intervals first, and then converted into HRV. As discussed previously, inferring HRV through RR intervals can lead to low accuracy due to error amplification. Moreover, as we show in this paper, pure signal processing may not be able to handle all types of PPG noises.

5.2 PPG HRV Monitoring with Machine Learning

While there are prior studies that employed ML models to estimate HRV, most of these studies focused on R peak/RR interval estimation, instead of predicting HRV directly. Everson et al. proposed a CNN-based encoder-decoder network to construct ECG waves from PPG waves and evaluated HRV based on the predicted wave [14]. They evaluated the model with the small ISPC dataset, which led to only one HRV estimation per recording. Similarly, Chiu et al. designed a CNN-based encoder-decoder with a sequence transformer network to generate ECG waves from PPG waves [11]. They evaluated the model with the UQVSD dataset and the BIDMC dataset - both are datasets from barely moving patients with low motion artifacts. Xu et al. classified PPG signals to systolic or diastolic phase using an RNN model with the assistance of an accelerometer [41]. They obtained RR intervals based on the classification results and evaluated the RR interval estimations. Wittenberg et al. compared a few CNN and GRU classification models for PPG R peaks detection [39]. They classified short PPG waves based on whether the first sample in it is an R peak. Maritsch et al. proposed a CNN model to predict the error of the RMSSD estimations from a smartwatch [22]. This work did not involve PPG signals. Alqaraawi et al. explored Bayesian learning to detect the PPG peaks with their collected data [3]. However, their monitoring length was only 5 or 8 minutes, and hence, only one HRV was provided, while our data lasted 2 hours. Choudhury et al. used a phone camera to capture signals of a human fingertip and extract RR intervals from the collected data with adaptive neural network (ANN) and SVM [12]. Instead of predicting HRV, most of the above studies mainly predicted/detected R peaks or RR intervals. However, as we show in Section 2.3, small errors in RR peak intervals can be magnified into large errors in HRV estimates. Therefore, in this work, we estimated HRV directly (represented by SDNN and RMSSD) from PPG signals.

Commercial wearable devices may also provide HRV estimations, such as Garmin smartwatches [16]. However, prior work has shown that smartwatch estimations may have high errors [22]. Due to their proprietary nature, we were not able

to rigorously evaluate commercial wearable devices. Therefore, we focused on comparing and analyzing existing research studies in our motivation and evaluation sections.

6 Conclusion

Photoplethysmography (PPG) sensors have been shown to be a good alternative for electrocardiographic (ECG) in Heart Rate Variability (HRV) monitoring. However, to be applied to practical and medical use, PPG HRV inference methods must be carefully designed. Prior work typically employed signal-processing-only or machine-learning-only methods to indirectly infer HRV from PPG signals, leading to low accuracy and large models. In this paper, we presented a compound and direct HRV inference method, which combines signal processing and machine learning to directly infer HRV. Evaluation results show that our method has errors as low as 3.5% with model sizes of a few hundred KBs, suggesting that our method can be applied in small embedded devices and potentially for medical uses.

Acknowledgment

This research was in part supported by the National Science Foundation (NSF), under grants, 2155096, 2221843, 2215359, 2309760, and 2317117. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the authors and do not necessarily reflect the view of the NSF. The authors would also like to thank the anonymous reviewers for their insightful comments

References

- [1] U Rajendra Acharya, K Paul Joseph, Natarajan Kannathal, Choo Min Lim, and Jasjit S Suri. 2006. Heart rate variability: a review. *Medical and biological engineering and computing* 44, 12 (2006), 1031–1051.
- [2] John Allen. 2007. Photoplethysmography and its application in clinical physiological measurement. *Physiological measurement* 28, 3 (2007).
- [3] Ahmed Alqaraawi, Ahmad Alwosheel, and Amr Alasaad. 2016. Heart rate variability estimation in photoplethysmography signals using Bayesian learning approach. *Healthcare technology letters* 3, 2 (2016).
- [4] Shikder Shafiul Bashar, Md Sazal Miah, AHM Zaididul Karim, Md Abdullah Al Mahmud, and Zahid Hasan. 2019. A machine learning approach for heart rate estimation from PPG signal using random forest regression algorithm. In *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*. IEEE, 1–5.
- [5] James Bergstra and Yoshua Bengio. 2012. Random search for hyperparameter optimization. *J. of machine learning research* 13, 2 (2012).
- [6] Tanmoy Bhowmik, Jishnu Dey, and Vijay Narayan Tiwari. 2017. A novel method for accurate estimation of HRV from smartwatch PPG signals. In *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE.
- [7] Dwaipayan Biswas, Luke Everson, Muqing Liu, et al. 2019. CorNET: Deep learning framework for PPG-based heart rate estimation and biometric identification in ambulant environment. *IEEE transactions on biomedical circuits and systems* 13, 2 (2019), 282–291.
- [8] Mike Blake. 2015. Development of a bluetooth 4.0 PPG sensor for use in Heart Rate Variability analysis. In *IEEE International Conference on Consumer Electronics (ICCE)*. IEEE, 301–304.

- [9] Denisse Castaneda, Aibhlinn Esparza, Mohammad Ghamari, Cinna Soltanpur, and Homer Nazeran. 2018. A review on wearable photoplethysmography sensors and their potential future applications in health care. *Journal of biosensors & bioelectronics* 4, 4 (2018), 195.
- [10] Xiangmao Chang, Gangkai Li, Guoliang Xing, Kun Zhu, and Linlin Tu. 2021. DeepHeart: A Deep Learning Approach for Accurate Heart Rate Estimation from PPG Signals. *ACM Transactions on Sensor Networks (TOSN)* 17, 2 (2021), 1–18.
- [11] Hong-Yu Chiu, Hong-Han Shuai, and Paul C-P Chao. 2020. Reconstructing QRS complex from PPG by transformed attentional neural networks. *IEEE Sensors Journal* 20, 20 (2020), 12374–12383.
- [12] Anirban Dutta Choudhury, Aishwarya Visvanathan, Rohan Banerjee, et al. 2013. Heartsense: estimating blood pressure and ecg from photoplethysmograph using smart phones. In *Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems*. 1–2.
- [13] Ashok Kumar Dohare, Vinod Kumar, and Ritesh Kumar. 2014. An efficient new method for the detection of QRS in electrocardiogram. *Computers & Electrical Engineering* 40, 5 (2014), 1717–1730.
- [14] Luke Everson, Dwaipayan Biswas, Bram-Ernst Verhoef, et al. 2019. BioTranslator: inferring R-peaks from ambulatory wrist-worn PPG signal. In *41st Annual International Conf. of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 4241–4245.
- [15] Jesse Fine, Kimberly L Branan, Andres J Rodriguez, et al. 2021. Sources of Inaccuracy in Photoplethysmography for Continuous Cardiovascular Monitoring. *Biosensors* 11, 4 (2021), 126.
- [16] Garmin. 2022. HRV Status. <https://www.garmin.com/en-US/garmin-technology/running-science/physiological-measurements/hrv-status/>.
- [17] Mohammad Ghamari, Cinna Soltanpur, S Cabrera, et al. 2016. Design and prototyping of a wristband-type wireless photoplethysmographic device for heart rate variability signal analysis. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 4967–4970.
- [18] Ary L. Goldberger, Zachary D. Goldberger, and Alexei Shvilkin. 2018. Chapter 3 - How to Make Basic ECG Measurements. In *Goldberger's Clinical Electrocardiography* (ninth edition ed.), Ary L. Goldberger, Zachary D. Goldberger, and Alexei Shvilkin (Eds.). Elsevier, 11–20.
- [19] Danilo Janković and Radovan Stojanović. 2017. Flexible system for HRV analysis using PPG signal. In *CMBEBIH 2017*. 705–712.
- [20] Robert E Kleiger, Phyllis K Stein, and J Thomas Bigger Jr. 2005. Heart rate variability: measurement and clinical utility. *Annals of Noninvasive Electrocardiology* 10, 1 (2005), 88–101.
- [21] Aboli N Londhe and Mithilesh Atulkar. 2019. Heart Rate Variability: A Methodological Survey. In *2019 International Conference on Intelligent Sustainable Systems (ICISS)*. IEEE, 57–63.
- [22] Martin Maritsch, Caterina Bérubé, Mathias Kraus, et al. 2019. Improving heart rate variability measurements from consumer smartwatches with machine learning. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*. 934–938.
- [23] William M Mendenhall and Terry L Sincich. 2016. *Statistics for Engineering and the Sciences*. CRC Press.
- [24] Task Force of the European Society of Cardiology the North American Society of Pacing. 1996. Heart Rate Variability: Standards of Measurement, Physiological Interpretation, and Clinical Use. *Circulation* 93, 5 (1996), 1043–1065.
- [25] Madhuri Panwar, Arvind Gautam, Dwaipayan Biswas, and Amit Acharyya. 2020. PP-Net: A Deep Learning Framework for PPG-Based Blood Pressure and Heart Rate Estimation. *IEEE Sensors Journal* 20, 17 (2020), 10000–10011.
- [26] F. Pedregosa, G. Varoquaux, A. Gramfort, et al. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.
- [27] Swapnil Puranik and Aldo W Morales. 2019. Heart rate estimation of PPG signals with simultaneous accelerometry using adaptive neural network filtering. *IEEE Trans. on Consumer Electronics* (2019).
- [28] Rama Reddy Rajanna, Sriraam Natarajan, and Prabhu Ravikala Vittal. 2018. An IoT Wi-Fi Connected Sensor For Real Time Heart Rate Variability Monitoring. In *2018 3rd International Conference on Circuits, Control, Communication and Computing (I4C)*. IEEE, 1–4.
- [29] Wala Saadeh, Shah Zaib Aslam, Aminah Hina, and Fakeha Asghar. 2018. A 0.5 V PPG-based heart rate and variability detection system. In *IEEE Biomedical Circuits and Systems Conference*. IEEE, 1–4.
- [30] Axel Schäfer and Jan Vagedes. 2013. How accurate is pulse rate variability as an estimate of heart rate variability?: A review on studies comparing photoplethysmographic technology with an electrocardiogram. *International journal of cardiology* 166, 1 (2013), 15–29.
- [31] RA Schneider and JP Costiloe. 1965. Relationship of Sinus Arrhythmia to Age and its Prognostic Significance in Ischemic Heart Disease. *Clin Res* 13 (1965), 219.
- [32] Lesia Semenova and Cynthia Rudin. 2019. A Study in Rashomon Curves and Volumes: A New Perspective on Generalization and Model Simplicity in Machine Learning. *CoRR* abs/1908.01755 (2019).
- [33] Murray R Spiegel. 2018. *Schaum's Outlines Statistics*. McGraw-Hill Education.
- [34] K Srinivas, L Ram Gopal Reddy, and R Srinivas. 2007. Estimation of heart rate variability from peripheral pulse wave using PPG sensor. In *3rd Kuala Lumpur International Conference on Biomedical Engineering 2006*. Springer, 325–328.
- [35] Rajendra G Sutar, Ashwin G Kothari, and AG Keskar. 2013. Development of an embedded system for real time Heart Rate Variability analysis. In *2013 13th International Symposium on Communications and Information Technologies (ISCIT)*. IEEE, 288–292.
- [36] DF Vysochanskij and Yu I Petunin. 1980. Justification of the 3σ rule for unimodal distributions. *Theory of Probability and Mathematical Statistics* 21, 25–36 (1980).
- [37] Lei Wang, Benny PL Lo, and Guang-Zhong Yang. 2007. Multichannel Reflective PPG Earpiece Sensor With Passive Motion Cancellation. *IEEE Trans. on Biomedical Circuits and Systems* 1, 4 (2007), 235–241.
- [38] Qing Wang, Zhao Wang, Xiaoming Dai, Shiwei Song, and Tianzhang Xing. 2019. S-HRVM: Smart Watch-based Heart Rate Variability Monitoring System.. In *Ewsn*. 178–183.
- [39] Th Wittenberg, R Koch, N Pfeiffer, et al. 2020. Evaluation of HRV estimation algorithms from PPG data using neural networks. *Current Directions in Biomedical Engineering* 6, 3 (2020), 505–509.
- [40] Borejda Xhyheri, Olivia Manfrini, Massimiliano Mazzolini, Carmine Pizzi, and Raffaele Bugiardi. 2012. Heart Rate Variability Today. *Progress in Cardiovascular Diseases* 55, 3 (2012), 321–331.
- [41] Ke Xu, Xinyu Jiang, Haoran Ren, Xiangyu Liu, and Wei Chen. 2019. Deep recurrent neural network for extracting pulse rate variability from photoplethysmography during strenuous physical exercise. In *IEEE Biomedical Circuits and Systems Conference (BioCAS)*. IEEE, 1–4.
- [42] Xiaohua Douglas Zhang. 2011. Illustration of SSMD, z score, SSMD*, z* score, and t statistic for hit selection in RNAi high-throughput screens. *Journal of biomolecular screening* 16, 7 (2011), 775–785.
- [43] Yuntong Zhang, Jingye Xu, Mimi Xie, Wei Wang, Keying Ye, Jing Wang, and Dakai Zhu. 2022. PPG-based Heart Rate Estimation with Efficient Sensor Sampling and Learning Models. In *2022 18th IEEE International Conference on Embedded Software and Systems (ICCESS)*. IEEE.
- [44] Zhilin Zhang. 2015. Photoplethysmography-based heart rate monitoring in physical activities via joint sparse spectrum reconstruction. *IEEE transactions on biomedical engineering* 62, 8 (2015), 1902–1910.
- [45] Zhilin Zhang, Zhouyue Pi, and Benyuan Liu. 2014. TROIKA: A general framework for heart rate monitoring using wrist-type photoplethysmographic signals during intensive physical exercise. *IEEE Transactions on biomedical engineering* 62, 2 (2014), 522–531.