WildPPG: A Real-World PPG Dataset of Long Continuous Recordings

Manuel Meier, Berken Utku Demirel, and Christian Holz
Department of Computer Science
ETH Zürich, Switzerland
firstname.lastname@inf.ethz.ch

Abstract

Reflective photoplethysmography (PPG) has become the default sensing technique in wearable devices to monitor cardiac activity via a person's heart rate (HR). However, PPG-based HR estimates can be substantially impacted by factors such as the wearer's activities, sensor placement and resulting motion artifacts, as well as environmental characteristics such as temperature and ambient light. These and other factors can significantly impact and decrease HR prediction reliability.

In this paper, we show that state-of-the-art HR estimation methods struggle when processing *representative* data from everyday activities in outdoor environments, likely because they rely on existing datasets that captured controlled conditions. We introduce a novel multimodal dataset and benchmark results for continuous PPG recordings during outdoor activities from 16 participants over 13.5 hours, captured from four wearable sensors, each worn at a different location on the body, totaling 216 hours. Our recordings include accelerometer, temperature, and altitude data, as well as a synchronized Lead I-based electrocardiogram for ground-truth HR references. Participants completed a round trip from Zurich to Jungfraujoch, a tall mountain in Switzerland over the course of one day. The trip included outdoor and indoor activities such as walking, hiking, stair climbing, eating, drinking, and resting at various temperatures and altitudes (up to 3,571 m above sea level) as well as using cars, trains, cable cars, and lifts for transport—all of which impacted participants' physiological dynamics. We also present a novel method that estimates HR values more robustly in such real-world scenarios than existing baselines.

Dataset & code for HR estimation: https://siplab.org/projects/WildPPG

1 Introduction

Today's wearable devices such as smartwatches and fitness trackers commonly monitor a person's cardiac health by continuously assessing their heart rate (HR). For estimating this metric, wearable devices predominantly use reflective photoplethysmography (PPG), which has become ubiquitous during ambulatory assessments due to its non-invasive nature and ease of use [1]. The HR measurements devices obtained from real-life conditions can supplement health assessments including exercise intensity, stress, fatigue, or sleep quality [2–4].

However, obtaining *reliable HR estimations* from PPG signals in real-world conditions and during real-world activities is challenging. Several external factors negatively impact accurate estimation, such as motion artifacts [5] that arise from a person's movements and performed activities [6, 7], sensor misplacement [8] or slippage during wear, and environmental conditions that change over time such as low temperatures [9] or high levels of ambient light [10]. Basic implementations of HR detection have been validated on comparably clean reference datasets, which do not adequately represent the variability and noise introduced by everyday activities and conditions. Even for simple

38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks.

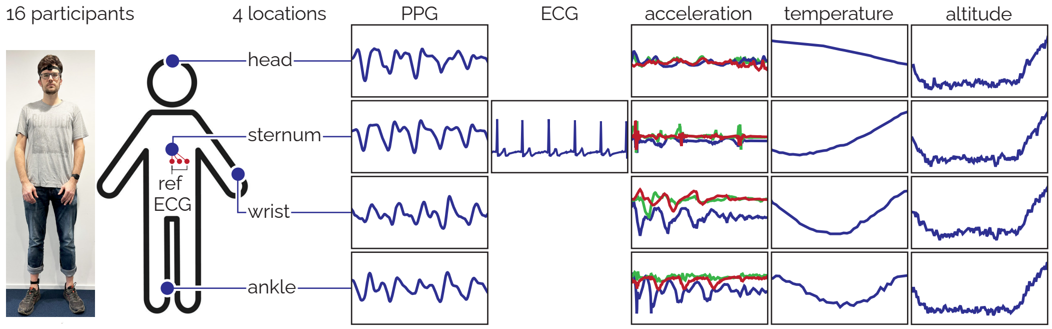


Figure 1: WildPPG comprises multi-modal signals from wearable devices at four sites on the body. Each device continuously recorded synchronized signals from a 3-channel reflective photoplethysmogram (red, green, infrared PPG), 3-axis inertial sensor (accelerometer), temperature, and barometric altitude sensor. For reference, the sternum device continuously recorded a Lead-I electrocardiogram (ECG) from body-mounted gel electrodes to provide ground-truth heart rate (HR) estimates.

activities such as walking and running [11], recent studies have provided evidence for the resulting detrimental effects on signal quality and, thus, the performance of HR tracking implementations.

To combat the effects of motions and activity on HR detection, prior work has proposed data-driven methods that learn to recover meaningful estimates under noisy conditions, including supervised methods [12–15], and rule-based techniques [5, 16–19]. While these and other methods estimate HR with high accuracy, their robustness to real-world effects is limited by the diversity of the datasets they were trained on. Because existing datasets were dominantly captured in controlled conditions, they may not adequately represent the noisy nature under which today's wearables obtain PPG measurements in daily life and "in the wild."

In this paper, we introduce WildPPG, a novel dataset of reflective PPG measurements from four locations across the wearer's body and reference recordings from an electrode-based Lead-I electrocardiogram (ECG) for HR estimation tasks from everyday activity outside controlled environments. Our dataset comprises continuous and synchronized recordings from 16 participants with an average duration of 13.5 hours during a round-trip to a tall mountain in Europe, including various daytime activities and modes of transport. As shown in Figure 1, our dataset additionally includes the continuous measurements from a 3-axis inertial sensor (accelerometer), barometric altitude sensor, and temperature sensor. Our data collection simultaneously recorded all these signals from four devices across the wearer's body (wrist, ankle, sternum, head), totaling 216 hours of synchronized multi-modal recordings that can enable a wide variety of future analyses.

We also contribute the implementation of multiple baseline methods for HR estimation and evaluate them on our dataset to identify weaknesses. Based on our analysis, we introduce a learning-based method that takes temperature as input in addition to the raw PPG signal and show that its temperature awareness allows it to produce more robust HR estimates.

Collectively, we make the following contributions in this dataset paper:

- a multi-modal real-world dataset of continuous, noisy, and synchronized PPG signals (raw red, green, infrared) from 4 body locations (forehead, sternum, wrist, ankle) and ECG recordings for reference, continuously collected from 16 participants over an average duration of 13.5 hours (minimum duration: 12.3 hours),
- a data collection study that covered everyday outdoor activities outside controlled (laboratory) conditions: During collection, participants completed a round-trip to a mountain station at over 3500 meters above sea level using various modes of transportation (car, train, cable car, elevator) and engaging in various outdoor and indoor activities (walking, hiking, stair climbing, eating, drinking, social interactions, restroom breaks, and resting in varying conditions), including environmental conditions of low and medium temperatures and various degrees of solar radiation,
- an accuracy analysis of baseline methods that estimate HR on WildPPG based on environmental characteristics. We also contribute a learning-based method for HR estimation that takes a raw PPG signal and temperature as input and show that it produces more robust estimates, and

- additionally recorded and synchronized modalities in our dataset as the basis for future algorithm developments, including 3-axis inertial measurements (accelerometer), barometric altitude, and temperature data from each of the 4 body locations (some exposed, some covered by clothing).

2 Related Work

2.1 PPG Datasets

Monitoring the cardiac activity of people using blood volume pulse signals, i.e., photoplethysmography, has a long history in continuous mobile health monitoring [26]. Robust HR estimation in real-world situations remains challenging due to the noisy nature of the PPG signals [5]. In order to develop, evaluate, and compare novel approaches, publicly available datasets are essential. However, existing public datasets are limited in terms of data size and applicability to wearable devices in real-world scenarios. For example the IEEE Signal Processing Challenge (SPC) 2015 dataset [16], includes 5-minute PPG recordings from 12 participants while the participants were controlled in the lab environments. This dataset is divided into two sets: a set with less motion artifacts from 12 participants (IEEE SPC12) and a set with heavy motions from 10 participants (IEEE SPC10). The BAMI dataset [22] was collected from 24 participants and includes resting, walking, and running activities. The experiments consisted of 2 minutes of walking as a warm-up, 3 minutes of running, 2 minutes of walking, 3 minutes of running, and 2 minutes of walking to cool down, all on a treadmill inside lab environments where even the speed was controlled (6.0 to 7.0 km/h). Similarly, the PPGMotion dataset contains walking, running, and cycling activities of 8 participants with an average recording duration of 17 minutes [25]. Meanwhile, the WESAD dataset [16] consists of approximately 1.6 hours of recordings for each of its 15 participants while they were seated and/or standing in a lab environment. It does not contain any physical activities but includes stress and amusement stimuli as well as meditation. All aforementioned datasets were recorded in controlled indoor conditions.

To record data that is more representative of real-world conditions, i.e., outside of lab environments, the DaLiA dataset [7] includes not only indoor but also outdoor activities such as cycling, walking, and driving. It contains recordings of 15 participants with a total of 26 hours of PPG data. Although this dataset is the largest among existing datasets, it is approximately eight times smaller than our presented dataset. Moreover, while the activities in DaLiA are controlled outside of lab environments, our experiments were conducted in completely free-living conditions, offering a more realistic representation of everyday activities. In other words, our dataset was collected "in the wild" — participants were not controlled and were free to move or engage in activities of their choice (within the framework of the experimental protocol) outside of the lab environment. It includes 216 hours of data, making it the largest dataset of its kind while including signals from multiple body locations with multiple wavelengths. Moreover, it is multi-modal. Covering the varying altitude-temperature changes, ensuring comprehensive and complete coverage of real-life conditions. Table 1 qualitatively compares WildPPG with the existing PPG datasets.

Table 1: Qualitative comparison of WildPPG with previous datasets for HR estimation.

Dataset (year)	**Collection methodology**	**Hours (total)**	**Body-worn locations** *	**Wavelengths (PPG)**	**Multi-modal**	**In-the-wild**** **dataset**	**Altitude/Temperature changes**
Ours (2024)	In the wild	216	{Head, Chest, Wrist, Ankle}	{Red, Green, Infra}	✔	✔	✔
Ear-PPG [20] (2023)	Very controlled	17	{Ear}	{Red, Green, Infra}	✔	✖	✖
Welltory [21] (2021)	Very controlled	1	{Wrist}	{Red, Green, Blue}	✖	✖	✖
DaLiA [7] (2019)	Controlled	36	{Wrist}	{Green}	✔	✔	✖
BAMI [22] (2019)	Very controlled	10	{Wrist}	{Red, Green, Infra}	✖	✖	✖
WESAD [23] (2018)	Controlled	25	{Finger}	{Green}	✔	✖	✖
BIDMC [24] (2017)	Very controlled	7	{Wrist}	—	✖	✖	✖
PPGMotion [25] (2017)	Very controlled	1.5	{Wrist}	{Green}	✖	✖	✖
IEEE SPC [16] (2015)	Very controlled	2	{Wrist}	{Red, Green}	✖	✖	✖

[*] *Body-worn locations* refer to the sites on the body where PPG signals are measured.
[**] *In-the-wild recording* is defined if the data were collected outside of (controlled) lab environments.

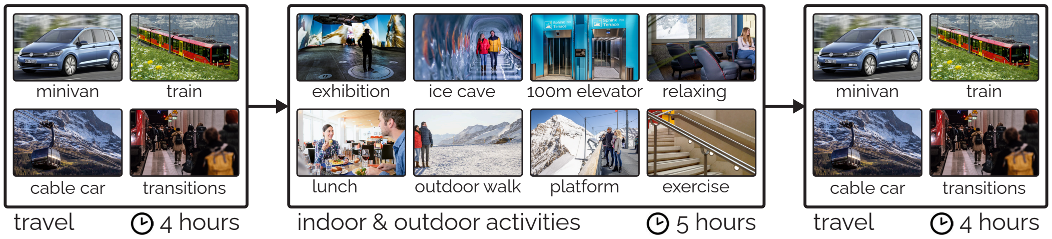


Figure 2: WildPPG participants engaged in multiple forms of travel as well as indoor and outdoor activities with changing environmental conditions. No strict study protocol was enforced and participants completed the activities at their own preferred speed.

2.2 HR estimation methods

Numerous efforts have investigated HR estimation from wearable devices using PPG signals during everyday motion and activity [18, 27–29], given its crucial role in mobile health monitoring [3]. Most of these studies employed Fourier transformation to observe how the frequency changes over time for estimating HR value [5, 27, 30]. While this approach works for less contaminated PPG signals, motion artifacts hinder the measurement of heart rate in spectral density. As a result, many approaches have been proposed to obtain the heart rate using the power spectral densities of PPG and accelerometer signals to differentiate the frequency of heartbeats from motion artifacts [14, 19, 27, 31, 32]. However, if the dominant frequencies in accelerometer signals overlap with the true heart rate, it becomes challenging to distinguish signals in the frequency domain [33]. In our previous work, we have shown that the combination of PPG signals from multiple body locations [17, 34] and the use of multiple PPG signals recorded at different wavelengths [13, 35] improves HR estimations, though motion artifacts often manifest across body locations and affect all PPG sensors. Our previous method BeliefPPG thus explicitly models uncertainties associated with PPG and IMU readings as well as estimated HR distribution, incorporating the statistical distribution of HR changes to refine estimates in a temporal context [36].

To enhance estimations based solely on PPG signals, recent works have proposed deep learning-based approaches [37, 38]. However, these models lack information about the degree of motion artifacts (MAs), leading to significant errors in heart rate estimation when inputs deviate from the training data distributions [31]. Even methods that combine input from a PPG sensor and an accelerometer struggle to learn the intricate relationship between motion artifacts and blood volume flow signals, resulting in unreliable heart rate predictions [18]. Thus, to address the limitations of existing PPG-based heart rate estimation methods, we introduce WildPPG which includes temperature as an additional modality alongside PPG recordings, resulting in improved heart rate estimation accuracy.

3 WildPPG Dataset

The purpose of WildPPG is to enable the development of more reliable PPG-based algorithms that improve the robustness of estimating HR as well as potential other cardiovascular metrics that are of interest on wearable devices such as smartwatches or fitness trackers. Our data recording methodology was designed to capture real-world data under less-than-optimal conditions during representative outdoor activities and everyday conditions. To achieve this, WildPPG contains long and uninterrupted recordings that our apparatus recorded from participants outside controlled environments or procedures. While our primary focus was on recording the signals crucial for HR estimation—PPG time series from a reflective green light-based sensing design across multiple body locations as well as chest ECG-based cardiac activity measurements for reference—our apparatus additionally recorded motion data from an inertial measurement unit at each body site, supplementary PPG signals from red and infrared reflections, as well as temperature readings and barometric altitude measurements throughout the recording for each participant.

Table 1 qualitatively compares our dataset's characteristics with commonly used datasets for HR prediction from blood volume pulse signals. We specifically focus on comparing our dataset with those collected under conditions similar to real-world conditions, where participants moved freely.

BIDMC is an exception, as it was collected with critically ill adults (a subset of the MIMIC-II dataset), and we include it because of its use in learning-based HR prediction in previous work (e.g., [12]). As shown in Table 1, our dataset is $\sim 8\times$ bigger than existing datasets for HR prediction and includes multi-modal data collected under real-world conditions. Below, we detail the experimental protocol and the design of our data capture apparatus for WildPPG.

3.1 Dataset Design

3.1.1 Experimental Protocol

Participants gathered in the morning to start the study. An experimenter outfitted each participant with four recording devices and ensured PPG and ECG signal quality (20 min). The participants then took a minivan from Zurich to Grindelwald (140 min) and transitioned to a cable car and train to Jungfraujoch railway station at 3460 m above sea level (80 min). Subsequently, they spent 5 hours in public places in and around the station and engaged in various activities. While the stay at the station included the same main activities for all participants, no strict study protocol was enforced in an effort to capture real-world data while participants completed the activities at their own speed of preference. Participants were encouraged to not constrain themselves due to the study protocol and engage in social interactions, take pictures, buy and consume snacks and beverages, take restroom breaks as needed, or sit down throughout the stay.

As shown in Figure 2, the main activities included walking through the museum and exhibition area including an ice cave with below-freezing temperatures (≈ 60 min), taking a 110-meter-high elevator and walking the remaining stairs up to the observatory and outside viewing platform (≈ 60 min), sitting down for lunch (≈ 60 min), walking through the snow-covered outside area (≈ 60 min), and resting inside (≈ 60 min). Spread out throughout the stay at Jungfraujoch, the participants additionally climbed and descended a set of stairs across four floors at maximum endurable pace on three to four occasions. After the stay at the station, participants took the train and cable car back to Grindelwald (80 min) and returned to Zurich on the minivan (140 min). Upon arrival, the experimenter removed the recording devices from each participant. Throughout the study, participants were instructed to remove the device on the wrist when washing their hands. All other devices were worn continuously.

3.1.2 Sensors

The signals collected in WildPPG were acquired using miniaturized, custom-built low-power wearable devices. As shown in Fig. 3, the devices are encased in a 3D-printed body and have an adjustable, stretchable strap with which they were attached at the forehead, sternum, ankle (supramalleolar), and wrist (dorsal). PPG measurements were obtained using an optical analog front-end at 128 Hz (MAX86141, Analog Devices) that connected to an optical module (SFH7072, ams-OSRAM) with a green (530 nm), a red (660 nm), and an infrared (950 nm) LED as well as a broadband photodiode (410 – 1100 nm), and an infrared-cut photodiode (402 – 694 nm). Green and red PPG were acquired in combination with the infrared-cut photodiode and infrared PPG with the broadband photodiode. Accelerometer data was acquired at a sampling rate of 200 Hz using a MEMS digital motion sensor (LIS2DH, STMicroelectronics). For ground truth, the sternum device additionally collected the Lead I ECG at 128 Hz through a biopotential sensor (MAX30003, Analog Devices) that connected to gel electrodes placed on the chest. Temperature and barometric altitude measurements were collected at a sampling rate of 10 Hz within the device (BME280, Bosch Sensortec).

Raw sensor data was continuously read from the sensors' FIFOs in batches by a System-on-a-Chip (DA14695, Dialog Semi), timestamped, stored in NAND memory (TH58CYG3S0HRAIJ, Kioxia Corp.), and downloaded after the recording was completed. Checksums were used to guarantee correct data download. Each device was powered by a CR2032 coin cell battery.

3.1.3 Synchronization

Since proper synchronization between sensors and devices is critical for this dataset, we collected the measurements "as synchronized as possible" on an electronics level and ensured proper trace alignment beyond the timing tolerances of the sensors. The most important measurements for this dataset, PPG and ECG, were triggered by an external clock derived from the real-time clock (RTC) of the System-on-a-Chip and thus perfectly synchronized with the system time of the device. This reduced the sample rate tolerance from $\pm 2\%$ of the internal sensor clock [39] by a factor of 100 to the

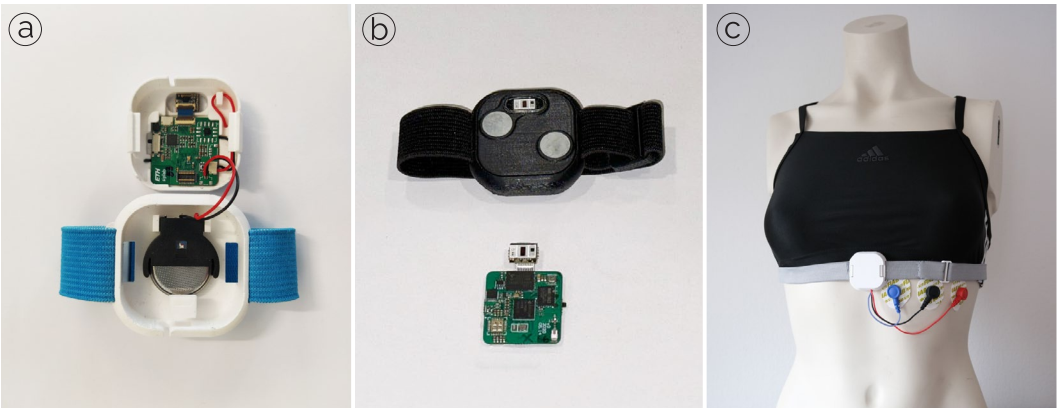


Figure 3: The wearable devices used for the data recording were custom-built and are centered around a SoC to read and store all synchronized sensor data in on-board flash memory. (a) Devices were powered by a CR2032 coin cell batteries (runtime: 18 hours) inside a 3D-printed case with (b) a flexible and adjustable strap. (c) The sternum unit additionally connected to 3 gel-electrodes to obtain a continuous ECG recording for reference.

tolerance of the used quartz (ECS-.327-7-16-C-TR, ECS Inc., ±20ppm tolerance [40]). Temperature and barometric pressure measurements were triggered based on the same RTC and therefore also perfectly synchronized to the PPG and ECG measurements. The accelerometer sampled based on its own internal clock. Deviations in sampling rate due to clock tolerances were corrected in post-processing based on the RTC timestamps of the FIFO readouts of the sensor.

This setup guarantees synchronization for the different sensing modalities acquired by a single device. However, measurements across devices are still participant to clock skew and drift effects – most notably when comparing PPG measurements from the wrist, ankle, and head devices with ground truth ECG measurements acquired on the chest. With the given quartz tolerance, although low, two devices may experience sampling shifts of up to 2 seconds across a 14-hour continuous recording session. Therefore, sensor data across devices was synchronized by aligning recorded signals offline following the approach described by Meier and Holz (33 ms accuracy) [41].

3.2 Recruitment and Recording

WildPPG's participants were 16 healthy adults, recruited on a voluntary basis without compensation beyond the coverage of all expenses incurred throughout the recording of the data and the visit of the mountain station. 7 female and 9 male participants took part, with ages ranging between 22 and 69 (mean age: 39). All participants' skin tones were within 1–3 on the Fitzpatrick scale [42]. All participants signed a consent form prior to participation. The experimenters conducted the data collection with groups of 3–5 participants, and the protocol was identical for all participant groups.

At the beginning of the recording, participants were introduced to the experimental protocol and the goal of the study. An experimenter was always with the participants throughout the recording. The recording devices worked without requiring user input, operating in a completely passive manner. The apparatus provided no feedback of any sort to participants either.

3.2.1 Risks

Participants were informed about the risk of developing symptoms of Acute Mountain Sickness (AMS) during the visit to Jungfraujoch station due to its altitude. Trains leaving to lower altitudes were available every 30 minutes in case a participant was feeling unwell. Additionally, a professional first aid station is available at Jungfraujoch alongside trained personnel and medical assistance if needed. The study protocol and risk assessment were approved by the ethics committee of ETH Zürich (EK 2022-N-44).

Table 2: The synchronized time series captured in WildPPG.

Time Series	**Description**	**depth** [bits]	**fps** [Hz]	**head**	**sternum**	**wrist**	**ankle**
ppg_g	Green PPG	19	128	×	×	×	×
ppg_ir	Infrared PPG	19	128	×	×	×	×
ppg_r	Red PPG	19	128	×	×	×	×
ecg	Lead I ECG	18	128		×		
accel_x	Accelerometer X-axis	10	128	×	×	×	×
accel_y	Accelerometer Y-axis	10	128	×	×	×	×
accel_z	Accelerometer Z-axis	10	128	×	×	×	×
altitude	Barometric altitude	23	0.5	×	×	×	×
temp	Temperature (inside case)	16	0.5	×	×	×	×

3.3 Dataset Composition

Participants' recordings are limited to the time they actually wore the sensor devices. PPG and ECG recordings are available as continuous recordings of raw sensor data. Accelerometer recordings were downsampled and synchronized to match the 128 Hz of the PPG and ECG signals. Device temperature measurements were averaged with an 8-second sliding window (2-second step size). The conversion from barometric pressure measurements to altitude above sea level was calculated following the conversion formula in the sensor's datasheet and a normed barometric pressure measurement from the same day acquired by a weather station at Jungfraujoch station run by the Swiss Federal Office of Meteorology and Climatology [43]. Barometric altitude measurements were averaged across all 4 devices worn by a participant and averaged with an 8-second sliding window and 2-second step size. Overall, WildPPG contains 216 hours of synchronized recordings. Accounting for the four separate device locations, this corresponds to 864 hours of PPG recordings at 3 different wavelengths. An overview of all captured data is shown in Table 2.

3.3.1 Ground Truth

As manual annotation of a dataset of this size is impractical and unreliable, the dataset contains Lead I-based ECG recordings for ground truth reference. For baseline analysis, we detected R-peaks in the ECG signal, using Pan-Tompkins [44]. To reduce the risk of corrupt ground truth values, peaks with inter-beat intervals (IBI) corresponding to HR values greater than 185 bpm or less than 35 bpm were removed, and for HR calculation, only IBIs that were part of a sequence of 4 IBIs for which $IBI_{min}/IBI_{max} < 0.75$ were considered. In data windows with less than two remaining IBI, no ground truth HR is computed and the window is omitted from baseline methods. Across the whole dataset, this applies to 2640 8-second windows (2.7% of total) mainly due to three participants with partially noisy ECG recordings which are responsible for 2164 of the rejected windows.

4 Baselines

We computed 6 heuristic and 5 supervised baseline algorithms as well as our own method on WildPPG. To improve comparability, all baselines were computed on data recorded by wrist-worn devices. All results are shown in Table 3 and the methods are described in more detail in the following two subsections.

4.1 Heuristic Methods

MSPTD is a peak detection method suited for PPG signals that computes a local maxima scalogram (LMS), a matrix that contains information about local maxima computed at different scales by comparing samples to neighbors at varying distances in the signal. The concept was first published by Scholkmann et al [45] and expanded upon by Bishop and Ercole by combining maxima detection with minima detection and improvements in algorithm efficiency [46]. The algorithm does not require any parameter tuning and works for any periodic or quasi-periodic signal. qppg is another peak detection algorithm built specifically for PPG signals and works by integrating the positive slope between diastolic and systolic points in the signal. Peaks are detected using adaptive thresholding [47]. HeartPy detects peaks in the PPG signal where it crosses its own moving average plus a variable offset. Multiple such offsets are dynamically tested and the one chosen which produces the most regularly spaced detected beats [48]. PWD is another PPG-specific method that relies on the detection of beats based on zero-crossings in the first derivative of the signal [49]. MSPTD, HeartPy, qppg, and PWD methods were computed using the MATLAB implementation by Charlton et al. which is released under GPL-3.0 open-source license [50]. We also included traditional Fourier transformation and autocorrelation functions, where both are used to detect periodicity [51].

4.2 Supervised Methods

For each supervised baseline, we follow the original implementation of the architectures. Additionally, we search for the best hyperparameters (learning rate and its scheduler, batch size, weight decay). Specifically, we compared our method with 1D ResNet [52] which is a modified version of ResNet architecture [53] for time series with 1D filters. DCL [54] is a widely used network that combines convolutional and long-short-term memory units for HR prediction from PPG signals [54,55]. Fully convolutional neural networks (FCN) which only include convolutional blocks with ReLU nonlinearities. LSTM [56] model which is designed to capture temporal dependencies in time series data through its memory cell architecture. Transformer [57] is a widely adopted deep learning model that employs self-attention mechanisms to capture long-range dependencies.

4.2.1 Proposed Method: Temp-ResNet

Considering the relation between the temperature of the measurement site and the SNR of PPG signals, we modified the 1D ResNet model by incorporating a small branch that inputs the temperature value of the segment to the overall architecture for HR estimation. Specifically, we use a multilayer perceptron with one hidden layer to obtain the representation for temperature $\mathbf{z}_{temp} = \mathbf{W}^{(2)}\sigma(\mathbf{W}^{(1)}\mathbf{x}_{temp})$ where σ and $\mathbf{x}_{temp}$ are a ReLU nonlinearity and temperature value of the specific segment, respectively. The weight matrices $\mathbf{W}^{(1)}$ and $\mathbf{W}^{(2)}$ are learnable parameters of the MLP, responsible for mapping the input temperature value to a higher-dimensional space. Initially, the temperature value $\mathbf{x}_{temp}$ is expanded to a 5-dimensional vector through the first layer, $\mathbf{W}^{(1)} \in \mathbb{R}^{5\times 1}$. This 5-dimensional representation is then further expanded to a 10-dimensional vector by the second layer, $\mathbf{W}^{(2)} \in \mathbb{R}^{5\times 10}$. Then, the obtained representation is concatenated with the extracted PPG

Table 3: Performance comparison of baselines with prior works in datasets

Method	WildPPG			SPC12*			DaLiA		
	MAE↓	RMSE↓	ρ↑	MAE↓	RMSE↓	ρ↑	MAE↓	RMSE↓	ρ↑
Heuristic									
FFT	17.62	28.68	12.11	14.83	25.81	35.15	34.98	47.13	2.72
Autocorrelation	28.43	35.23	-3.06	41.16	49.81	3.945	30.59	39.09	9.18
HeartPy	18.49	30.15	14.7	19.89	24.81	20.19	12.25	18.15	74.97
MSPTD	13.30	21.19	21.37	20.35	25.31	18.22	17.20	25.11	53.95
PWD	11.72	19.57	25.19	21.16	26.32	18.75	28.26	37.48	29.51
qppgfast	19.32	29.94	21.27	17.97	21.99	32.72	13.31	20.35	66.58
Self-supervised									
SimCLR	15.75±1.81	19.81±1.17	8.12±2.51	12.42±0.05	20.96±0.30	60.41±0.52	12.01±0.14	19.46±0.14	58.31±0.39
NNCLR	14.46±0.13	18.43±0.05	11.54±0.56	13.14±0.49	18.86±0.49	69.82±0.06	12.94±0.31	20.02±0.49	51.12±2.54
BYOL	14.11±0.12	18.42±0.15	12.50±0.50	18.71±0.93	25.01±1.50	48.82±4.36	11.67±0.32	17.57±0.23	63.96±0.97
TS-TCC	12.64±0.03	17.72±0.04	18.71±0.41	11.56±0.41	18.04±0.66	68.38±1.41	8.12±0.30	14.89±0.21	67.13±0.53
TS2Vec	10.55±0.37	16.52±0.39	26.31±0.25	9.75±0.08	17.82±0.43	75.43±0.33	10.83±0.13	17.89±0.19	60.10±0.62
VICReg	15.37±0.60	19.20±0.33	9.30±1.57	13.17±0.82	20.38±1.27	59.76±4.16	14.90±0.16	21.94±0.11	45.38±0.07
Barlow Twins	16.14±0.92	20.21±0.90	9.13±1.20	13.22±0.34	20.42±0.88	64.51±4.01	18.26±0.57	23.41±0.29	23.42±7.30
Unsupervised									
UPD	23.12±1.10	25.10±0.56	7.39±1.83	9.30±0.10	16.50±0.20	77.60±0.43	27.41±4.73	31.26±4.55	18.12±3.86
Supervised									
1D ResNet	8.62±0.06	15.00±0.06	42.41±0.24	**5.50**±0.29	**11.16**±0.64	**84.86**±0.51	4.47±0.03	10.03±0.10	85.87±0.25
DCL	8.64±0.01	14.73±0.05	41.39±0.11	16.50±1.42	21.61±1.56	19.90±1.10	26.34±1.10	25.53±4.2	80.57±0.09
FCN	9.79±0.23	15.62±0.26	35.68±0.94	27.55±0.55	31.18±0.81	40.41±1.81	6.55±0.28	11.24±0.32	83.21±0.04
LSTM	9.28±0.22	15.14±0.11	36.96±1.21	18.66±3.49	25.02±3.45	56.23±4.84	5.30±0.05	11.10±0.15	78.99±0.44
Transformer	10.06±0.16	16.23±0.20	32.07±1.60	22.00±0.23	27.30±0.29	52.86±0.10	7.84±0.11	15.06±0.17	68.62±0.53
Temp-ResNet	**8.37**±0.02	**14.72**±0.01	**45.50**±0.14	—	—	—	**4.40**±0.02	**9.87**±0.11	**85.90**±0.17

* The dataset has no temperature value from the measurement site.

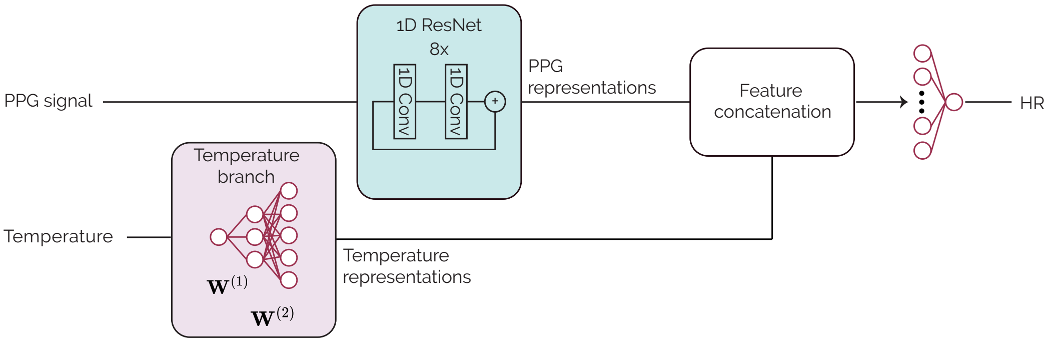


Figure 4: Architecture of Temp-ResNet. The ResNet includes batch normalization and ReLU activations after each convolution.

representations and fed to the final linear layer for HR estimation. The architecture of Temp-ResNet is shown in Figure 4.

5 Limitations

Regarding skin pigmentation, all participants were within 1–3 on the Fitzpatrick scale, and no participants rated themselves 4–6 (darker skin tones). Considering the more challenging nature of PPG measurements among people with darker skin tones [58, 59], it will be important to broaden data collection in future efforts. More generally, including 16 participants is a limitation of our dataset, and including more would allow covering as many participant-specific effects as possible. Lastly, as described in Section 3.3.1, 2.7% of the total dataset is affected by noisy ground truth measurements.

6 Ethical Considerations

Working with real-world physiological and activity data from humans requires ethical considerations. All participants were informed regarding safety and privacy both verbally and in writing. They signed a consent form that they received prior to the study for careful reading. No personal information is released with the dataset, the data is fully anonymous and does not contain personally identifiable information or offensive content. To the best of our knowledge, no potential harm or malicious use is possible using the WildPPG dataset and the provided methods. The study was approved by the ethics committee of ETH Zürich.

7 Conclusion

We have presented WildPPG, a real-world dataset of long and continuous multimodal motion and physiology recordings for PPG-based HR analysis. Our experimental protocol included a wide variety of activities and environmental conditions including indoor and outdoor activities at different ambient temperatures, altitude levels, and light conditions. Participants moved, acted, and interacted naturally throughout the day and recording without being bound to a strict study procedure—for the purpose of capturing representative data that resembles the measurements taken by wearable devices in *everyday* use. Besides PPG measurements at the three common red, green, and infrared wavelengths and ECG ground-truth references, WildPPG includes acceleration, temperature, and barometric altitude measurements to suit a wide range of analysis methods and tasks. Our study recorded and synchronized measurements from multiple body locations to support exploring future estimation techniques that exceed the analysis of signals from wrist-worn devices most common on smartwatches today and to capture cardiac activity more holistically and robustly in the future.

To the best of our knowledge, WildPPG is the largest available dataset of its kind and can enable learning pipelines that benefit from longer time series during training, additional modalities, and real-world data ("in the wild"). Beyond the dataset, WildPPG includes a range of baseline algorithms—heuristic as well as supervised—and our novel method that leverages temperature readings for improved HR estimation. These methods can serve as a starting point for future work and benchmarks and our intention is to enable future work to better tailor methods to real-world situations and applications in less-than-optimal conditions. This could include multi-modal processing or analyses of additional cardiovascular metrics that are of interest in wearable devices.

Acknowledgments and Disclosure of Funding

This work has not received any third-party funding and the authors declare no conflicts of interest.

References

[1] K. Bayoumy, M. Gaber, A. Elshafeey, O. Mhaimeed, E. H. Dineen, F. A. Marvel, S. S. Martin, E. D. Muse, M. P. Turakhia, K. G. Tarakji, and M. B. Elshazly, “Smart wearable devices in cardiovascular care: where we are and how to move forward,” *Nature Reviews Cardiology*, vol. 18, no. 8, pp. 581–599, Aug. 2021. [Online]. Available: https://www.nature.com/articles/s41569-021-00522-7

[2] S. Huhn, M. Axt, H.-C. Gunga, M. A. Maggioni, S. Munga, D. Obor, A. Sié, V. Boudo, A. Bunker, R. Sauerborn, T. Bärnighausen, and S. Barteit, “The impact of wearable technologies in health research: Scoping review,” *JMIR Mhealth Uhealth*, vol. 10, no. 1, p. e34384, Jan 2022.

[3] A. Hughes, M. M. H. Shandhi, H. Master, J. Dunn, and E. Brittain, “Wearable devices in cardiovascular medicine,” *Circulation Research*, vol. 132, no. 5, pp. 652–670, 2023. [Online]. Available: https://www.ahajournals.org/doi/abs/10.1161/CIRCRESAHA.122.322389

[4] P. K. Stein and Y. Pu, “Heart rate variability, sleep and sleep disorders,” *Sleep Medicine Reviews*, vol. 16, no. 1, pp. 47–66, 2012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1087079211000293

[5] Z. Zhang, Z. Pi, and B. Liu, “Troika: A general framework for heart rate monitoring using wrist-type photoplethysmographic signals during intensive physical exercise,” *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 2, pp. 522–531, 2015.

[6] F. Sarhaddi, K. Kazemi, I. Azimi, R. Cao, H. Niela-Vilén, A. Axelin, P. Liljeberg, and A. M. Rahmani, “A comprehensive accuracy assessment of samsung smartwatch heart rate and heart rate variability,” *PLOS ONE*, vol. 17, no. 12, pp. 1–19, 12 2022. [Online]. Available: https://doi.org/10.1371/journal.pone.0268361

[7] A. Reiss, I. Indlekofer, P. Schmidt, and K. Van Laerhoven, “Deep ppg: Large-scale heart rate estimation with convolutional neural networks,” *Sensors*, vol. 19, no. 14, 2019. [Online]. Available: https://www.mdpi.com/1424-8220/19/14/3079

[8] S. Lee, H. Shin, and C. Hahm, “Effective PPG sensor placement for reflected red and green light, and infrared wristband-type photoplethysmography,” in *2016 18th International Conference on Advanced Communication Technology (ICACT)*, Jan. 2016, pp. 556–558. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/7423470

[9] M. Khan, C. G. Pretty, A. C. Amies, R. Elliott, G. M. Shaw, and J. G. Chase, “Investigating the Effects of Temperature on Photoplethysmography,” *IFAC-PapersOnLine*, vol. 48, no. 20, pp. 360–365, Jan. 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2405896315020571

[10] J. Kim, T. Lee, J. Kim, and H. Ko, “Ambient light cancellation in photoplethysmogram application using alternating sampling and charge redistribution technique,” in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Aug. 2015, pp. 6441–6444, iSSN: 1558-4615. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/7319867

[11] S. Benedetto, C. Caldato, E. Bazzan, D. C. Greenwood, V. Pensabene, and P. Actis, “Assessment of the fitbit charge 2 for monitoring heart rate,” *PLOS ONE*, vol. 13, no. 2, pp. 1–10, 02 2018. [Online]. Available: https://doi.org/10.1371/journal.pone.0192691

[12] P. Sarkar and A. Etemad, “Cardiogan: Attentive generative adversarial network with dual discriminators for synthesis of ecg from ppg,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 1, pp. 488–496, May 2021. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/16126

[13] M. Meier, B. U. Demirel, and C. Holz, "Tri-spectral ppg: Robust reflective photoplethysmography by fusing multiple wavelengths for cardiac monitoring," in *2024 IEEE 20th International Conference on Body Sensor Networks (BSN)*. IEEE, 2024, pp. 1–4.

[14] S. S. Chowdhury, R. Hyder, M. S. B. Hafiz, and M. A. Haque, "Real-time robust heart rate estimation from wrist-type ppg signals using multiple reference adaptive noise cancellation," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 2, pp. 450–459, 2018.

[15] L. G. Rocha, G. Paim, D. Biswas, S. Bampi, F. Catthoor, C. Van Hoof, and N. Van Helleputte, "Lstm-only model for low-complexity hr estimation from wrist ppg," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2021, pp. 1068–1071.

[16] Z. Zhang, "Photoplethysmography-based heart rate monitoring in physical activities via joint sparse spectrum reconstruction," *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 8, pp. 1902–1910, 2015.

[17] M. Meier and C. Holz, "Robust heart rate detection via multi-site photoplethysmography," in *2024 46th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2024, pp. 1–4.

[18] A. Temko, "Accurate heart rate monitoring during physical exercises using ppg," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 9, pp. 2016–2024, 2017.

[19] S. M. A. Salehizadeh, D. Dao, J. Bolkhovsky, C. Cho, Y. Mendelson, and K. H. Chon, "A novel time-varying spectral filtering algorithm for reconstruction of motion artifact corrupted heart rate signals during intense physical activities using a wearable photoplethysmogram sensor," *Sensors*, vol. 16, no. 1, 2016. [Online]. Available: https://www.mdpi.com/1424-8220/16/1/10

[20] A. Montanari, A. Ferlini, A. N. Balaji, C. Mascolo, and F. Kawsar, "EarSet: A Multi-Modal Dataset for Studying the Impact of Head and Facial Movements on In-Ear PPG Signals," *Scientific Data*, vol. 10, no. 1, p. 850, Dec. 2023. [Online]. Available: https://www.nature.com/articles/s41597-023-02762-3

[21] A. Neshitov, K. Tyapochkin, E. Smorodnikova, and P. Pravdin, "Wavelet analysis and self-similarity of photoplethysmography signals for hrv estimation and quality assessment," *Sensors*, vol. 21, no. 20, 2021. [Online]. Available: https://www.mdpi.com/1424-8220/21/20/6798

[22] H. Lee, H. Chung, and J. Lee, "Motion artifact cancellation in wearable photoplethysmography using gyroscope," *IEEE Sensors Journal*, vol. 19, no. 3, pp. 1166–1175, 2019.

[23] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. Van Laerhoven, "Introducing wesad, a multimodal dataset for wearable stress and affect detection," in *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, ser. ICMI '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 400–408. [Online]. Available: https://doi.org/10.1145/3242969.3242985

[24] M. A. F. Pimentel, A. E. W. Johnson, P. H. Charlton, D. Birrenkott, P. J. Watkinson, L. Tarassenko, and D. A. Clifton, "Toward a robust estimation of respiratory rate from pulse oximeters," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 8, pp. 1914–1923, 2017.

[25] D. Jarchi and A. J. Casson, "Description of a database containing wrist ppg signals recorded during physical exercise with both accelerometer and gyroscope measures of motion," *Data*, vol. 2, no. 1, 2017. [Online]. Available: https://www.mdpi.com/2306-5729/2/1/1

[26] *M-Health*. [Online]. Available: https://link.springer.com/book/10.1007/b137697

[27] Q. Xie, Q. Zhang, G. Wang, and Y. Lian, "Combining adaptive filter and phase vocoder for heart rate monitoring using photoplethysmography during physical exercise," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2018, pp. 3568–3571.

[28] C. Holz and E. J. Wang, "Glabella: Continuously sensing blood pressure behavior using an unobtrusive wearable device," vol. 1, 2017.

[29] A. Carek and C. Holz, "Naptics: convenient and continuous blood pressure monitoring during sleep," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 3, pp. 1–22, 2018.

[30] M. Zhou and N. Selvaraj, “Heart rate monitoring using sparse spectral curve tracing,” in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2020, pp. 5347–5352.

[31] A. Reiss, I. Indlekofer, P. Schmidt, and K. Van Laerhoven, “Deep ppg: Large-scale heart rate estimation with convolutional neural networks,” *Sensors*, vol. 19, no. 14, 2019. [Online]. Available: https://www.mdpi.com/1424-8220/19/14/3079

[32] T. Schäck, M. Muma, and A. M. Zoubir, “Computationally efficient heart rate estimation during physical exercise using photoplethysmographic signals,” in *2017 25th European Signal Processing Conference (EUSIPCO)*, 2017, pp. 2478–2481.

[33] D. Pollreisz and N. TaheriNejad, “Detection and Removal of Motion Artifacts in PPG Signals,” *Mobile Networks and Applications*, vol. 27, no. 2, pp. 728–738, Apr. 2022. [Online]. Available: https://doi.org/10.1007/s11036-019-01323-6

[34] M. Meier and C. Holz, “Assessing the accuracy of photoplethysmography for wearable heart rate monitoring based on body location and body motion in uncontrolled outdoor environments,” *Current Issues in Sport Science (CISS)*, vol. 9, no. 2, Feb. 2024.

[35] ——, “Impact of optical wavelength on the reliability of photoplethysmography-based heart rate measurements outside of controlled laboratory environments,” *Current Issues in Sport Science (CISS)*, vol. 9, no. 2, Feb. 2024.

[36] V. Bieri, P. Streli, B. U. Demirel, and C. Holz, “BeliefPPG: uncertainty-aware heart rate estimation from ppg signals via belief propagation,” in *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, ser. UAI ’23. PMLR, 2023, pp. 173–183.

[37] A. Shyam, V. Ravichandran, S. Preejith, J. Joseph, and M. Sivaprakasam, “Ppgnet: Deep network for device independent heart rate estimation from photoplethysmogram,” in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2019, pp. 1899–1902.

[38] L. G. Rocha, D. Biswas, B.-E. Verhoef, S. Bampi, C. Van Hoof, M. Konijnenburg, M. Verhelst, and N. Van Helleputte, “Binary cornet: Accelerator for hr estimation from wrist-ppg,” *IEEE Transactions on Biomedical Circuits and Systems*, vol. 14, no. 4, pp. 715–726, 2020.

[39] “MAX86141 Datasheet and Product Info | Analog Devices.” [Online]. Available: https://www.analog.com/en/products/max86141.html

[40] A. Raphael, “ECS-.327-7-16-C-TR.” [Online]. Available: https://ecsxtal.com/products/crystals/surface-mount-crystals/ecs-327-7-16-c-tr/

[41] M. Meier and C. Holz, “BMAR: Barometric and Motion-based Alignment and Refinement for Offline Signal Synchronization across Devices,” *Proc. ACM IMWUT*, vol. 7, no. 2, pp. 69:1–69:21, Jun. 2023.

[42] T. B. Fitzpatrick, “The validity and practicality of sun-reactive skin types i through vi,” *Archives of dermatology*, vol. 124, no. 6, pp. 869–871, 1988.

[43] “Homepage - MeteoSwiss.” [Online]. Available: https://www.meteoswiss.admin.ch/

[44] J. Pan and W. J. Tompkins, “A Real-Time QRS Detection Algorithm,” *IEEE Transactions on Biomedical Engineering*, vol. BME-32, no. 3, pp. 230–236, Mar. 1985, conference Name: IEEE Transactions on Biomedical Engineering. [Online]. Available: https://ieeexplore.ieee.org/document/4122029

[45] F. Scholkmann, J. Boss, and M. Wolf, “An Efficient Algorithm for Automatic Peak Detection in Noisy Periodic and Quasi-Periodic Signals,” *Algorithms*, vol. 5, no. 4, pp. 588–603, Dec. 2012, number: 4 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/1999-4893/5/4/588

[46] S. M. Bishop and A. Ercole, “Multi-Scale Peak and Trough Detection Optimised for Periodic and Quasi-Periodic Neuroscience Data,” in *Intracranial Pressure & Neuromonitoring XVI*, T. Heldt, Ed. Cham: Springer International Publishing, 2018, pp. 189–195.

[47] A. N. Vest, G. D. Poian, Q. Li, C. Liu, S. Nemati, A. J. Shah, and G. D. Clifford, “An open source benchmarked toolbox for cardiovascular waveform and interval analysis,” *Physiological Measurement*, vol. 39, no. 10, p. 105004, Oct. 2018, publisher: IOP Publishing. [Online]. Available: https://dx.doi.org/10.1088/1361-6579/aae021

[48] P. van Gent, H. Farah, N. van Nes, and B. van Arem, "HeartPy: A novel heart rate algorithm for the analysis of noisy signals," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 66, pp. 368–378, Oct. 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1369847818306740

[49] B. N. Li, M. C. Dong, and M. I. Vai, "On an automatic delineator for arterial blood pressure waveforms," *Biomedical Signal Processing and Control*, vol. 5, no. 1, pp. 76–81, Jan. 2010. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1746809409000470

[50] P. H. Charlton, K. Kotzen, E. Mejía-Mejía, P. J. Aston, K. Budidha, J. Mant, C. Pettit, J. A. Behar, and P. A. Kyriacou, "Detecting beats in the photoplethysmogram: benchmarking open-source algorithms," *Physiological Measurement*, vol. 43, no. 8, p. 085007, Aug. 2022, publisher: IOP Publishing. [Online]. Available: https://dx.doi.org/10.1088/1361-6579/ac826d

[51] L. Saul and J. Allen, "Periodic component analysis: An eigenvalue method for representing periodic structure in speech," in *Advances in Neural Information Processing Systems*, T. Leen, T. Dietterich, and V. Tresp, Eds., vol. 13. MIT Press, 2000. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2000/file/f22e4747da1aa27e363d86d40ff442fe-Paper.pdf

[52] S. Hong, Y. Xu, A. Khare, S. Priambada, K. Maher, A. Aljiffry, J. Sun, and A. Tumanov, "Holmes: Health online model ensemble serving for deep learning models in intensive care units," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 1614–1624.

[53] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015.

[54] H. Qian, T. Tian, and C. Miao, "What makes good contrastive learning on small-scale wearable-based tasks?" in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, ser. KDD '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 3761–3771. [Online]. Available: https://doi.org/10.1145/3534678.3539134

[55] D. Biswas, L. Everson, M. Liu, M. Panwar, B.-E. Verhoef, S. Patki, C. H. Kim, A. Acharyya, C. Van Hoof, M. Konijnenburg, and N. Van Helleputte, "Cornet: Deep learning framework for ppg-based heart rate estimation and biometric identification in ambulant environment," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 13, no. 2, pp. 282–291, 2019.

[56] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, 1997.

[57] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[58] B. A. Fallow, T. Tarumi, and H. Tanaka, "Influence of skin type and wavelength on light wave reflectance," *Journal of clinical monitoring and computing*, vol. 27, pp. 313–317, 2013.

[59] P. E. Bickler, J. R. Feiner, and J. W. Severinghaus, "Effects of skin pigmentation on pulse oximeter accuracy at low saturation," *The Journal of the American Society of Anesthesiologists*, vol. 102, no. 4, pp. 715–719, 2005.

[60] "CC BY-NC-SA 4.0 Deed | Attribution-NonCommercial-ShareAlike 4.0 International | Creative Commons." [Online]. Available: https://creativecommons.org/licenses/by-nc-sa/4.0/

[61] "GNU General Public License version 3," Oct. 2007. [Online]. Available: https://opensource.org/license/gpl-3-0

[62] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research. PMLR, 2015. [Online]. Available: https://proceedings.mlr.press/v37/ioffe15.html

[63] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 56, pp. 1929–1958, 2014. [Online]. Available: http://jmlr.org/papers/v15/srivastava14a.html

[64] B. U. Demirel and C. Holz, "An unsupervised approach for periodic source detection in time series," 2024. [Online]. Available: https://arxiv.org/abs/2406.00566

[65] B. U. Demirel and C. , Holz, “Finding order in chaos: A novel data augmentation method for time series in contrastive learning,” in *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36. Curran Associates, Inc., 2023, pp. 30 750–30 783. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/file/61c2c6338033da68885e0226881cbe71-Paper-Conference.pdf

[66] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1412.6980

[67] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in *Proceedings of the 37th International Conference on Machine Learning*, ser. ICML’20. JMLR.org, 2020.

[68] D. Dwibedi, Y. Aytar, J. Tompson, P. Sermanet, and A. Zisserman, “With a little help from my friends: Nearest-neighbor contrastive learning of visual representations,” in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. Los Alamitos, CA, USA: IEEE Computer Society, oct 2021, pp. 9568–9577. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/ICCV48922.2021.00945

[69] E. Eldele, M. Ragab, Z. Chen, M. Wu, C. K. Kwoh, X. Li, and C. Guan, “Time-series representation learning via temporal and contextual contrasting,” in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, 2021, pp. 2352–2359.

[70] Z. Yue, Y. Wang, J. Duan, T. Yang, C. Huang, Y. Tong, and B. Xu, “Ts2vec: Towards universal representation of time series,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 8, pp. 8980–8987, Jun. 2022. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/20881

[71] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, “Barlow twins: Self-supervised learning via redundancy reduction,” in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 18–24 Jul 2021, pp. 12 310–12 320. [Online]. Available: https://proceedings.mlr.press/v139/zbontar21a.html

[72] S. K. Longmore, G. Y. Lui, G. Naik, P. P. Breen, B. Jalaludin, and G. D. Gargiulo, “A Comparison of Reflective Photoplethysmography for Detection of Heart Rate, Blood Oxygen Saturation, and Respiration Rate at Various Anatomical Locations,” *Sensors*, vol. 19, no. 8, p. 1874, Jan. 2019, number: 8 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/1424-8220/19/8/1874

[73] D. Ray, T. Collins, S. I. Woolley, and P. V. S. Ponnapalli, “A Review of Wearable Multi-Wavelength Photoplethysmography,” *IEEE Reviews in Biomedical Engineering*, vol. 16, pp. 136–151, 2023, conference Name: IEEE Reviews in Biomedical Engineering. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9582790

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] see Section 5
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] see Section 6
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Section B.2
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] Section B.2
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] 4.1
   (b) Did you mention the license of the assets? [Yes] 4.1
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] 6
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] 6
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A] Not applicable.
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [Yes] 3.2.1
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [Yes] 3.2

Appendix

A Dataset Accessibility

We provide WildPPG time series in the MATLAB .mat file format. Code is provided in Matlab .m file format and Python .py source code. The python source code contains functions to load the .mat files into equivalent python structures. All data, code, and the project website are hosted on ETH Zurich servers with long-running maintenance intended for long-term availability. The dataset uses a Creative Commons CC BY-NC-SA 4.0 license [60] while the code is released under GPL-3 [61].

B Experimental Settings

This section provides more details on the implementations of the learning-based baseline methods and the dataset we tested them on.

B.1 Datasets

The datasets used in this study are both widely employed in related work about HR estimation from PPG signals.

IEEE SPC This competition provided a training dataset of 12 participants (SPC12) and a test dataset of 10 participants [5]. The IEEE SPC dataset overall has 22 recordings of 22 participants (ages 18 – 58) performing three different activities. Each recording contains data sampled at 125 Hz from a 3-axis accelerometer and 2-channel pulse oximeter sensor (green LEDs). All these recordings were captured from the wearable device placed on the wrist of each individual. Additionally, a chest ECG provides ground-truth HR estimation. During our experiments, we used the PPG channels only. We use leave-one-out-cross-validation similar to the previous setups.

DaLiA PPG signals from the DaLia dataset were recorded at a sampling rate of 64 Hz. We used the whole dataset of 15 participants while following the leave-one-out-cross-validation. 10% of the training data is randomly split for validation, i.e., early stopping and saving the best model.

All PPG datasets are standardized as follows. Initially, a fourth-order Butterworth bandpass filter with a frequency range of 0.5–4 Hz is applied to the PPG signals. Subsequently, a sliding window of 8 seconds with 2-second shifts is employed for segmentation, followed by z-score normalization of each segment. Lastly, the signal is resampled to a frequency of 25 Hz for each segment.

B.2 Supervised Baseline Methods

We train the models using a batch size of 128 with a learning rate of $5e-4$. The validation loss is employed to save the optimal model and halt training to prevent overfitting. If no improvements are observed for 15 consecutive epochs, we reduce the learning rate by a factor of 10. Our experiments utilized NVIDIA GeForce RTX 4090 GPUs and involved training with three random seeds across all datasets, resulting in approximately 480 GPU hours, including ablation.

1D ResNet is a modified version of ResNet for time series with 1D filters of size five across eight residual blocks. The initial block consists of 32 filters, incrementing by a factor of two for every two subsequent residual blocks. Also, batch normalization [62] is applied after each convolutional block. We apply a pooling operation after each residual with a stride of 2 while using Dropout [63] with 0.5 after each activation. Finally, a global average pooling is used before the final linear layer.

DCL is a combination of convolutional blocks with the long-short term memory units [56] (LSTMs) and is widely used for time series [54, 55] as it considers the temporal relationship. Specifically, the DCL architecture has four convolutional layers with 5×1 size of 64 kernels. Then, the output is fed into the 2-layer LSTM with 128 units. The last time step of the LSTM is fed to a linear layer.

FCN We also implemented a fully convolutional neural network with a 3-layer followed by ReLU activation and MaxPooling after each convolutional layer, similar to the implementation in [54].

Dropout with 0.5 is applied after the first convolutional layer. We set the kernel and padding size to 8 and 4, respectively for each convolutional layer. The number of kernels for each convolutional layer is set to 32 for the first one and 64 for the rest.

LSTM We employed a unit of two layers of LSTM [56] without any feature extraction, i.e., the raw PPG data is fed to the model. The features from the final LSTM are fed to a linear layer for HR estimation.

Transformer We also added transformers with positional encodings as a baseline model. Since the attention mechanism [57] is used to capture dependencies between any points within the input sequence, irrespective of their temporal distance, we used it for time series data, similar to [54]. Specifically, we used linear layers with a stack of four identical blocks. The linear layer converts the input data to embedding vectors of 128. A token of size 128 is added to the embedded input as the representation vector. Each block is made up of a multi-head self-attention layer and a fully connected feed-forward layer with residual connections around them.

B.3 Self-Supervised Baseline Methods

For our self-supervised learning experiments, we follow the same implementation setup as previous works [54, 64, 65] for self-supervised learning in time series. Specifically, we use a combination of convolutional with LSTM-based network, which is widely used in heart rate estimation from PPG signals [55, 65], as backbones for the encoder $f_\theta(.)$ where the projector is two fully connected layers. During pre-training, we use InfoNCE (for contrastive learning-based methods) as the loss function, which is optimized using Adam [66] with a learning rate of 0.003. We train the models with a batch size of 256 for 120 epochs and decay the learning rate using the cosine decay. After pre-training, we train a single linear layer classifier on features extracted from the frozen pre-trained network, i.e., linear probing.

SimCLR SimCLR [67] introduces a contrastive learning framework for self-supervised visual representation learning. The method relies on maximizing agreement between differently augmented views of the same image via a contrastive loss in the latent space. We follow the previous implementations of SimCLR for time series [54] and PPG signals [65].

NNCLR We follow a similar setup to SimCLR by applying two separate data augmentations, and then we use nearest neighbors in the learned representation space as the positive in contrastive losses [68]. The maximum size of the support set equals 1024.

BYOL For the BYOL implementation, the exponential moving average parameter is set to 0.996 where the projector size is set to 128. We set the learning rate to 0.03 similar to other SSL techniques. Following the original implementation, we use a weight decay parameter of $1.5e-6$.

TS-TCC We follow the same architecture implementation with the losses, i.e., contextual and temporal contrasting. TS-TCC [69] proposed applying two separate augmentations, one augmentation is weak (jitter-and-scale) and the other is strong (permutation-and-jitter). The authors also change the strength of the permutation window from dataset to dataset. In our experiments, we follow the previous work [64] for PPG-based augmentation.

TS2Vec TS2Vec [70] is an SSL method specifically designed for time series based on contrastive (instance and temporal-wise) learning in a hierarchical way over augmented context views where the context is generated by applying timestamp masking and random cropping on the input time series. Following the original framework, we use a dilated CNN architecture with a depth of 10 and a hidden size of 64, which has a similar number of parameters to the architectures used by other SSL methods. The batch size is set to 256, and the number of epochs to 120, consistent with other SSL techniques.

VICReg We follow the original implementation and set the coefficients for each loss term to 25 (λ), 25 (μ), and 1 (ν), corresponding to the invariance, variance, and covariance terms, respectively. We have not performed an additional hyper parameter search as these values are also set by previous on this application [64].

Barlow Twins Barlow Twins [71] presents a function to avoid collapse for SSL by measuring the cross-correlation matrix between the outputs of two identical networks fed with augmented versions of a sample, and making it as close to the identity matrix as possible. This causes the embedding vectors of augmented versions of a sample to be similar while minimizing the redundancy between the components of these vectors. Following the original implementation, we applied batch normalization to the extracted embeddings and set the hyperparameter λ coefficient (in Equation 1) to 0.005.

$$\mathcal{L} = \sum_{i}(1 - C_{ii})^2 + \lambda \sum_{i} \sum_{j \neq i} C_{ij}^2, \tag{1}$$

where C is the cross-correlation matrix computed between the two sets of normalized embeddings.

B.4 Unsupervised Baseline Method

We also consider one of the unsupervised learning based method, unsupervised periodicity detection (UPD), as an additional baseline model.

UPD UPD [64] presented two novel regularizers for detecting periodic patterns in time series data without using labels or specific augmentations. We follow the original implementation `https://github.com/eth-siplab/Unsupervised_Periodicity_Detection`.

Since this method operates without supervision, it requires relatively clean samples to learn periodic representations. Consequently, its performance is quite low on datasets with noisy samples, such as WildPPG and DaLiA datasets.

C Additional Experiments

Here, we present additional experiments we conducted beyond our main experiments, which focused on estimating the heart rate from green PPG recordings on the wrist as it is the most widely used measurement site and PPG wavelength for wearable devices [11].

C.1 Evaluation Across Multiple Body Locations and PPG Wavelengths

We investigated the performance of the models when we integrate PPG signals from the different measurement sites of the body as well as PPG signals recorded at different wavelengths. Table 4 presents the results for the baseline and the modified ResNet architecture performances.

Table 4: Performance comparison of baselines with prior works in datasets

Location	WildPPG		
Method	MAE↓	RMSE↓	ρ↑
Only Wrist			
1D ResNet	8.62±0.06	15.00±0.06	42.41±0.24
Temp-ResNet	8.37±0.02	14.72±0.01	45.50±0.14
Only Chest			
1D ResNet	8.55±0.05	14.95±0.07	44.30±0.20
Temp-ResNet	8.35±0.03	14.68±0.02	46.48±0.10
Wrist and Chest			
1D ResNet	8.40±0.10	14.85±0.10	44.40±0.20
Temp-ResNet	**8.28**±0.03	**14.67**±0.02	**47.55**±0.13
All			
1D ResNet	8.60±0.05	14.93±0.10	43.27±0.21
Temp-ResNet	8.33±0.05	14.70±0.05	46.50±0.10

Wavelength	WildPPG		
Method	MAE↓	RMSE↓	ρ↑
Green			
1D ResNet	8.62±0.06	15.00±0.06	42.41±0.24
Temp-ResNet	8.37±0.02	**14.72**±0.01	45.50±0.14
Red			
1D ResNet	10.25±0.14	17.63±0.21	35.21±0.23
Temp-ResNet	9.97±0.10	16.41±0.12	36.45±0.10
Infrared			
1D ResNet	8.97±0.12	15.93±0.12	36.33±0.16
Temp-ResNet	8.72±0.10	15.88±0.11	37.66±0.09
All			
1D ResNet	8.59±0.03	14.97±0.05	43.21±0.13
Temp-ResNet	**8.35**±0.04	14.68±0.10	**46.38**±0.10

From these results, we observe that incorporating additional measurement sites for HR prediction generally reduces the error, despite increasing the number of model parameters. Furthermore, results suggest that the wrist measurement site has the lowest signal quality compared to other sites. This outcome is expected, as the wrist is particularly susceptible to motion artifacts and exposure to cold, and the results are consistent with prior work [72]. Similarly, the signal quality of alternative wavelengths is lesser compared to green PPG which is also consistent with prior work [73].

C.2 Cross-Dataset Evaluation

We evaluated model performance using a cross-dataset approach, specifically employing a leave-one-dataset-out scheme. In this setup, two datasets were used for training, while the third dataset was reserved for testing. Results are summarized in Table 5.

Table 5: Performance comparison of baselines in cross-dataset evaluation (one dataset left for evaluation and others used for training)

Method	Test on WildPPG			Test on SPC12			Test on DaLiA		
	MAE↓	RMSE↓	ρ↑	MAE↓	RMSE↓	ρ↑	MAE↓	RMSE↓	ρ↑
Supervision									
1D ResNet	16.32±0.26	27.33±0.33	12.33±0.24	**11.25**±1.78	**18.37**±0.88	**67.36**±0.78	17.23±2.28	22.10±0.13	67.14±0.15
DCL	19.43±2.12	29.32±2.43	10.68±2.59	15.45±2.34	26.32±2.09	55.43±4.75	20.90±0.55	25.31±1.54	60.21±2.56
FCN	16.55±1.23	26.77±0.26	11.53±1.56	14.22±2.45	24.58±1.96	50.28±1.50	24.13±1.13	24.56±2.31	65.32±2.56
LSTM	16.79±1.30	28.57±2.43	8.97±1.10	27.86±2.21	28.11±4.10	42.11±3.95	21.59±1.10	23.47±1.67	63.42±1.30
Transformer	17.25±1.45	30.98±2.59	6.72±3.66	28.45±1.98	30.38±3.59	30.85±2.05	20.31±2.56	20.17±2.17	55.21±2.17

C.3 Additional sensor modality

In Table 6 we present the experimental results of integrating acceleration with PPG as an additional modality.

Table 6: Performance comparison of baseline models across datasets when integrating accelerometer data with PPG as an additional modality

Method	Test on WildPPG			Test on SPC12			Test on DaLiA		
	MAE↓	RMSE↓	ρ↑	MAE↓	RMSE↓	ρ↑	MAE↓	RMSE↓	ρ↑
Supervision									
1D ResNet	8.22±0.09	14.34±0.07	44.33±0.22	**5.33**±0.18	**11.08**±0.67	**85.02**±0.47	4.43±0.10	09.88±0.13	85.96±0.20
DCL	8.69±0.03	15.03±0.09	42.01±0.07	15.30±1.20	20.98±1.33	20.78±1.26	4.88±0.12	11.73±0.07	79.12±0.12
FCN	9.65±0.20	15.16±0.22	37.42±0.71	25.95±0.37	27.87±0.12	45.13±1.80	6.12±0.23	10.92±0.28	84.04±0.03
LSTM	8.72±0.20	14.11±0.10	38.61±1.10	17.41±2.30	21.13±2.57	60.10±3.12	5.07±0.08	10.94±0.20	80.01±0.30
Transformer	9.98±0.15	17.10±0.23	31.73±1.08	21.92±0.20	25.10±0.18	50.97±0.21	7.17±0.15	14.98±0.20	68.97±0.66
Temp-ResNet	**8.13**±0.03	**13.95**±0.02	**47.48**±0.10	—	—	—	**4.27**±0.03	**9.13**±0.15	**87.31**±0.19

C.4 Baseline Results on the BIDMC Dataset

Considering the wide use of the BIDMC dataset (recorded in controlled environments), we provide baseline results of the baseline methods listed in Section B.2 for this dataset in Table 7.

Table 7: Performance comparison of baselines when tested on the BIDMC dataset

Method	BIDMC		
	MAE↓	RMSE↓	ρ↑
Heuristic			
FFT	4.41	9.74	33.42
Supervision			
1D ResNet	**3.62**±0.22	**5.23**±0.39	85.43±0.20
DCL	4.18±0.31	5.89±0.65	83.34±1.37
FCN	4.01±0.27	5.41±0.47	84.57±1.14
LSTM	5.73±1.79	10.86±0.96	29.37±1.58
Transformer	7.71±1.40	11.03±1.35	19.21±2.20

D Impact of Temperature and Motion

The modified architecture Temp-ResNet adds the device's case temperature (measured within the encasing) as an additional input modality to improve results. Figure 5 shows the HR error of 1D ResNet based on PPG from the wrist in relation to case temperature and motion. The correlations of error and temperature/motion are supported by the investigation of signal-to-noise ratio (SNR): We have computed the SNR of the PPG signal by relating the power P_{HR} of the ground truth HR in the frequency spectrum of the PPG signal to the power $P_{noise} = P_{tot} - P_{HR}$ of the other frequency components in the relevant window of possible heart rates:

$$SNR = 10 \cdot log \left(\frac{P_{HR}}{P_{tot} - P_{HR}} \right) dB$$

While it is well understood that motion artifacts have a negative impact on PPG signal quality [5], the presented results show that the temperature within the wearable has an inverse correlation to PPG signal quality as well.

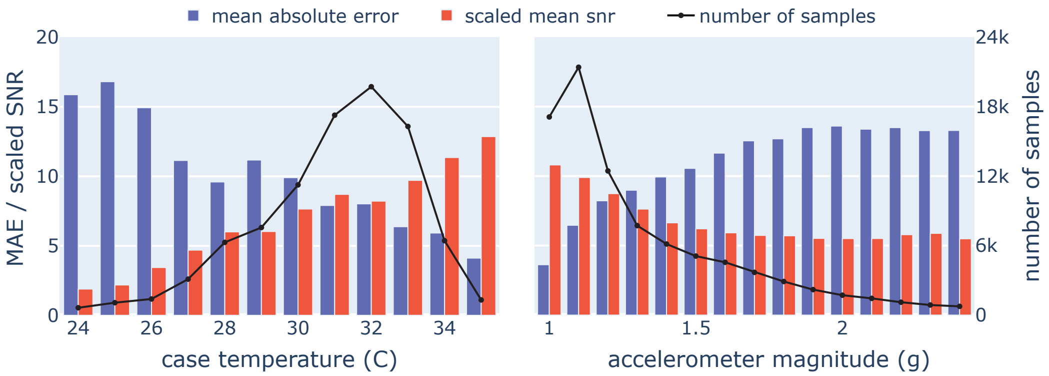


Figure 5: The error (blue) of 1D ResNet based on PPG from the wrist, dependent on the device temperature (left), and device motion (right). Inversely related is the computed SNR of the PPG signal (red).

E PPG Device

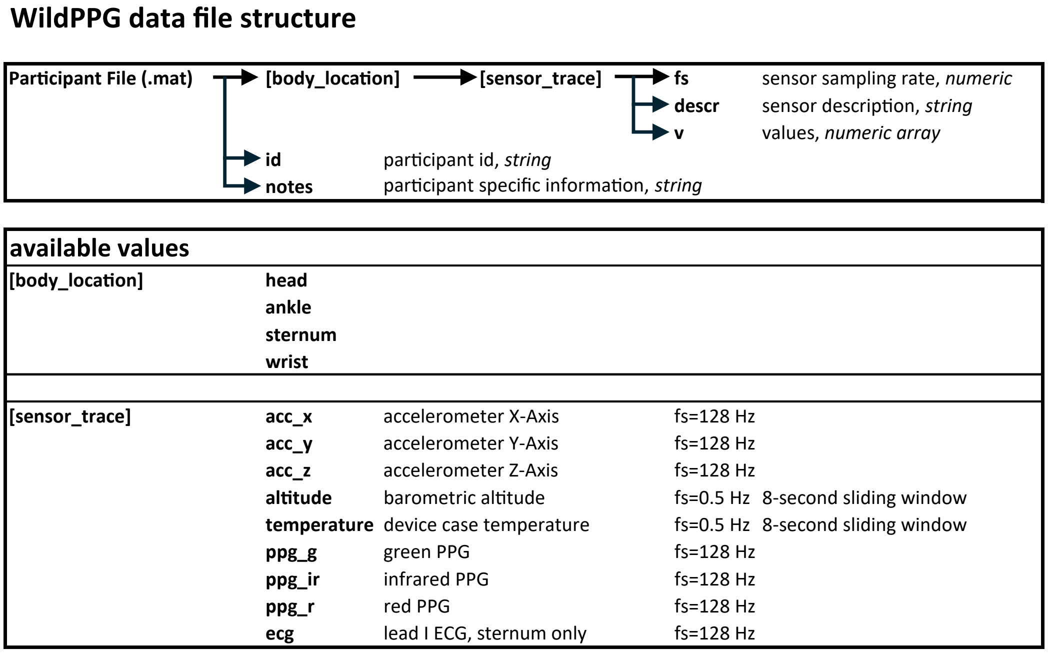
The schematics and production data (CAD drawing, BOM, Gerber/NCDrill, pick&place) of the PPG device used in this project are available on the project website. The data is acquired using an off-the-shelf PPG LED module (SFH7072, OSRAM) and a commonly used dedicated PPG acquisition chip (MAX86141, Analog Devices). Table 8 shows the register configuration of the MAX86141 PPG chip as used during the data acquisition.

Table 8: Register configurations of the MAX86141 PPG chip.

Register Name	Reg Addr	Value	Description
MAX86141_PPG_CONFIG_1	0x11	0x2B	ALC enable, 117us integration time, 16uA ADC range
MAX86141_PPG_CONFIG_2	0x12	0x70	128 Hz SR (ext. clock), no averaging
MAX86141_PPG_CONFIG_3	0x13	0xC0	12uS LED settle time
MAX86141_PHOTO_DIODE_BIAS	0x15	0x11	Low capacity PD
MAX86141_LED_RANGE_1	0x2A	0x00	LED driver range 31mA
MAX86141_LED1_PA	0x23	0x12	IR LED current 2.16mA
MAX86141_LED2_PA	0x24	0x0C	R LED current 1.44mA
MAX86141_LED3_PA	0x25	0x0C	G LED current 1.44mA
MAX86141_INTERRUPT_ENABLE_1	0x02	0x00	Disable all Interrupts
MAX86141_LED_SEQ_REG1	0x20	0x21	Set LED sequence (1 slot each for IR R G)
MAX86141_LED_SEQ_REG2	0x21	0x03	Set LED sequence (1 slot each for IR R G)
MAX86141_LED_SEQ_REG3	0x22	0x00	Set LED sequence (1 slot each for IR R G)

F File Structure

The file structure of the WildPPG dataset is shown in Figure 6



WildPPG data file structure

Participant File (.mat) → **[body_location]** → **[sensor_trace]** →	**fs**	sensor sampling rate, *numeric*
	descr	sensor description, *string*
	v	values, *numeric array*
Participant File (.mat) →	**id**	participant id, *string*
	notes	participant specific information, *string*

available values

[body_location]	**head**			
	ankle			
	sternum			
	wrist			
[sensor_trace]	**acc_x**	accelerometer X-Axis	fs=128 Hz	
	acc_y	accelerometer Y-Axis	fs=128 Hz	
	acc_z	accelerometer Z-Axis	fs=128 Hz	
	altitude	barometric altitude	fs=0.5 Hz	8-second sliding window
	temperature	device case temperature	fs=0.5 Hz	8-second sliding window
	ppg_g	green PPG	fs=128 Hz	
	ppg_ir	infrared PPG	fs=128 Hz	
	ppg_r	red PPG	fs=128 Hz	
	ecg	lead I ECG, sternum only	fs=128 Hz	

Figure 6: For each participant in WildPPG, one .mat Matlab data file is provided with the structure as illustrated.

G Author Statement

The authors bear all responsibility in case of violation of rights.