

WildPPG - Datasheet

Manuel Meier, Berken Utku Demirel, Christian Holz
Department of Computer Science, ETH Zurich
{manuel.meier, berken.demirel, christian.holz}@inf.ethz.ch

This document is based on *Datasheets for Datasets* by Gebre *et al.* [2].

MOTIVATION

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

WildPPG addresses the need for a large, real-world (outside controlled environments) dataset of PPG data, associated ECG ground truth as well as additional modalities. It also includes data from different body locations to explore PPG measurements beyond its most common use in wrist-worn devices such as smartwatches. Beyond that, PPG measurements are provided at three wavelengths for additional possible analysis.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

This dataset was created by SIPlab at the Department of Computer Science of ETH Zurich

What support was needed to make this dataset? (e.g. who funded the creation of the dataset? If there is an associated grant, provide the name of the grantor and the grant name and number, or if it was supported by a company or government agency, give those details.)
This project did not receive any third-party funding.

Any other comments?

-

COMPOSITION

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

The instances are physiological recordings from human participants in a study. For each participant, there are multiple recordings from different devices which are all time synchronized. There are no interactions between the recordings of different participants.

How many instances are there in total (of each type, if appropriate)?

There are 16 participants with recordings from 4 devices each.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

This dataset contains all possible instances.

What data does each instance consist of? “Raw” data (e.g., unprocessed text or images) or features? In either case, please provide a description.

Time-series data from physiological sensors.

Is there a label or target associated with each instance? If so, please provide a description.

Each device instance is labeled with a body location where it was worn.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

No.

Are relationships between individual instances made explicit (e.g., users’ movie ratings, social network links)? If so, please describe how these relationships are made explicit.

Multiple devices are grouped into instances of participants. There are no complex relationships between instances beyond that.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

Splitting by participant and/or device location on the body is the recommended choice to assess how well an algorithm

generalizes across a population or different signal sources.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

All recordings in the dataset are affected by measurement noise of the sensors that collected them. For 3 participants, the ECG recordings, which may be used for ground truth purposes, are temporarily affected by strong noise due to electrical contact issues at the electrodes.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset is self-contained.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.

No.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

No.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

Yes.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

No.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

No, the data is anonymous.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic

data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

Yes, the data contains physiological data which may allow conclusions about the cardiovascular health of a participant. However, this is mitigated by the fact that the data is anonymous and participants explicitly agreed to the recording and publication of the data.

Any other comments?

-

COLLECTION

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

All data was directly observable / measured by sensors.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created. Finally, list when the dataset was first published.

Each instance was recorded over the course of a day and the timeframe of the data is identical. Different instances were recorded over the course of 4 months.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

The data was collected using custom-built wearable devices with off-the-shelf sensor chips: 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 a 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 using a combined

sensor (BME280, Bosch Sensortec) which is located within the device. The wearable devices were validated using medical laboratory equipment (Biopac Systems Inc.)

What was the resource cost of collecting the data? (e.g. what were the required computational resources, and the associated financial costs, and energy consumption - estimate the carbon footprint. See Strubell *et al.*[4] for approaches in this area.)

The study protocol included travel by car which totaled 1400km across the whole data collection which is roughly equivalent to 200kg of CO₂ emissions and 600 USD in financial cost. Additional transportation was fully electric (cable car, train), powered by energy sources with minimal emissions. Besides this, no significant computational resources were used. The sensing hardware is reusable with a negligible amortization of production cost/emissions associated with this data collection.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

-

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

The authors of this dataset were the experimenters who accompanied the participants of the study. Compensation was covered by their regular work contracts. Participants were recruited on a voluntary basis and received no compensation beyond the coverage of all expenses incurred during the data recording.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

Yes, the study was approved by the ethics board of ETH Zurich (EK 2022-N-44). This included a review with regards to physical safety of the participants, data security and other ethical considerations.

Does the dataset relate to people? If not, you may skip the remainder of the questions in this section.
Yes.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

No questions were part of the data recording beyond demographic data which was assessed directly.

Were the individuals in question notified about the data collection? If so, please describe (or show with

screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.
Yes, they explicitly chose to join the data recording.

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

Yes, all participants signed a consent form which informed them about the data collection. They agreed to the recording and publication of the data.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate)

Generally, participants can revoke their consent at any time by explicit written request through e-mail to the authors. However, participants agreed to the exclusion of the anonymous data published as part of this dataset from this mechanism after the point of its publication.

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

Because the data is published anonymized, no such analysis was conducted.

Any other comments?

-

PREPROCESSING / CLEANING / LABELING

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

Yes, data across devices was synchronized following the approach by Meier and Holz [3] to achieve sample-by-sample synchronization across devices. Temperature and altitude data were averaged in a moving window approach with a window length of 8 seconds and step size of 2 seconds.

Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the “raw” data.

Yes, the full raw data is stored on internal servers. Access may be requested from the authors directly through email.

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

No, the software is not publicly available.

Any other comments?

-

USES

Has the dataset been used for any tasks already? If so, please provide a description.

Currently, there is no published work relating to the dataset.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

Currently, no such repository exists.

What (other) tasks could the dataset be used for?

Due to its multimodal nature, the dataset may be used for a variety of physiological analyses, activity monitoring, and other methods that relate to the provided data.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

No.

Are there tasks for which the dataset should not be used? If so, please provide a description.

The dataset must not be used for the diagnosis of conditions relating to the health condition of any participant. No analysis must be conducted that may lead to curtailing the anonymity of the data.

Any other comments?

-

DISTRIBUTION

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

Yes, the dataset will be publicly available under Creative

Commons license.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

The dataset will be accessible through a website hosted by ETH Zürich with long-term maintenance.

When will the dataset be distributed?

When it is accepted for publication at a peer-reviewed venue.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

The dataset is released under Creative Commons BY-SA 4.0 license [1].

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

No.

Any other comments?

-

MAINTENANCE

Who is supporting/hosting/maintaining the dataset?

The dataset is hosted on ETH Zurich servers with long-running maintenance intended for long-term availability.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

All authors may be contacted through email as listed in the header of the document.

Is there an erratum? If so, please provide a link or other access point.

No.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will

be communicated to users (e.g., mailing list, GitHub)?

If the authors discover or are made aware of issues with the dataset, it will be updated and users will be informed through the project website including a changelog.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

No.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

Old versions or scripts to roll back the current version of the dataset will be available through the project website.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

The dataset cannot be contributed to by outside contributors. The authors may extend the dataset in the future but do so in a way that does not compromise the ability to work with the original dataset only.

Any other comments?

-

REFERENCES

- [1] CC BY-SA 4.0 Deed | Attribution-ShareAlike 4.0 International | Creative Commons.
- [2] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Dauméé III, and Kate Crawford. Datasheets for Datasets. *arXiv:1803.09010 [cs]*, January 2020.
- [3] Manuel Meier and Christian Holz. BMAR: Barometric and Motion-based Alignment and Refinement for Offline Signal Synchronization across Devices. *Proc. ACM IMWUT*, 7(2):69:1–69:21, June 2023.
- [4] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and Policy Considerations for Deep Learning in NLP. *arXiv:1906.02243 [cs]*, June 2019.