VitalLens: Take A Vital Selfie

Philipp V. Rouast
Rouast Labs
philipp@rouast.com

Abstract

This report introduces VitalLens, an app that estimates vital signs such as heart rate and respiration rate from selfie video in real time. The estimation engine of VitalLens is a computer vision model trained on a diverse dataset of video and gold-standard ground truth physiological sensor data.

1 Introduction

Video of the human face and upper body contains signals with information about the subject's physiological state [6]. Given appropriate circumstances, these signals can be strong enough to estimate vital signs such as heart rate and respiratory rate. This process is known as Remote Photoplethysmography (rPPG); many rPPG approaches have been proposed, including handcrafted algorithms and ones learned from empirical data. Depending on the intended application, we can measure advantages and drawbacks of rPPG approaches in the accuracy, inference speed, and robustness regarding various factors impacting estimation performance.

Paragraph on G [6], CHROM [2], and POS [7].

Paragraph on DeepPhys [1] and MTTS-CAN [3].

Acknowledge other models we don't compare because lack focus on real-time.

[INSERT FIGURE WITH SAMPLE PREDICTIONS FROM VALIDATION SET]

Section summarizing the remainder of the report.

2 Model Architecture

The estimation engine built into VitalLens is a computer vision model trained on a diverse dataset of video and gold-standard ground truth physiological data. The model architecture is broadly based on the *EfficientNetV2* [4] model family, enhanced with several improvements in architecture and model optimization to enable efficient training and inference in the rPPG domain.

This report does not contain any further detail about the architecture of the model or training methodology - we reserve this for future publication.

3 Datasets

VitalLens is trained on the *PROSIT* and *Vital Videos Ghana* datasets. The vast majority of training data comes from PROSIT. For evaluation, we use the entire *Vital Videos Medium* dataset as well as the test sets of the *PROSIT* and *Vital Videos Ghana* datasets.

3.1 PROSIT

PROSIT (Physiological Recordings Of Subjects using Imaging Technology) is our in-house dataset collected in Australia for practical rPPG applications.

Participant recruitment and session protocol. As part of our goal of creating a diverse rPPG dataset for practical applications, we recruit participants and collect data at various locations such as residential homes, offices, libraries, and clubs. Each potential participant was required to go through an informed consent process in accordance with Australian privacy laws prior to participation. During the session, participants are asked to complete tasks on a hand-held iPad while sensor data is being recorded. Participants are not explicitly asked to remain stationary, which results in varying amounts of participant movement. For some sessions, the camera is on a tripod and thereby stationary, while for other sessions the camera is fixed to the iPad and thereby not necessarily stationary.

Sensors and collected data. The time-synchronized sensor array used for PROSIT consists of a video camera, electrocardiogram (ECG), pulse oximetry, blood pressure monitor, and an ambient light sensor. This yields a rich set of data including video, ECG, PPG, SpO2, respiration, blood pressure, and ambient luminance. We also collect age, gender, height, and skin type metadata according to the Fitzpatrick scale.

Pre-processing. We pre-process and split each session into small chunks of 5-20 seconds with valid video and signals. As part of this step, we also calculate summary vitals for each chunk from the continuous signals, and extract further metadata such as the amount of participant movement and illumination variation.

Dataset size and split. Development of PROSIT is ongoing. As of the writing of this report, it comprises 157 unique participants across 173 recording sessions in 45 different locations. This results in a total of 9,767 chunks or 27.8 hours of data.

Table 1: PROSIT Dataset Size

Split	# Participants	# Chunks	Time
Train	114	6,765	19.4 h
Valid	23	1,599	4.5 h
Test	20	1,403	3.9 h
Total	157	9,767	27.8 h

Each participant is randomly assigned to be part of either the *training*, *validation*, or *test* set to ensure that all participants seen during validation and test are previously unseen by the model.

3.2 Vital Videos Medium

Vital Videos is a large, diverse dataset for rPPG collected in Belgium [5]. It is the largest dataset we are aware of that is available for research without academic affiliation. We use a slightly extended version of the medium instantiation ("VV-Medium"), which consists of 296 participants. Both camera and participants are stationary in this dataset.

Pre-processing. We pre-process VV-Medium using the same steps applied to PROSIT. As part of this step, we create small chunks, calculate summary vitals, and extract further metadata.

Table 2: VV-Medium Dataset Size

Split	# Participants	# Chunks	Time
Test	289	1,108	4.6 h

The entirety of VV-Medium is used to test the capabilities of VitalLens.

3.3 Vital Videos Ghana

Vital Videos Ghana is a new dataset from the authors of Vital Videos aiming to address the insufficient share of participants with skin types 5 and 6 usually found in rPPG datasets. It was collected in

Ghana, with the majority of participants having skin type 5 or 6. We use a small instantiation ("VV-Ghana-Small"), which consists of 129 participants. Both camera and participants are stationary in this dataset.

Pre-processing. We pre-process VV-Ghana-Small using the same steps applied to PROSIT. As part of this step, we create small chunks, calculate summary vitals, and extract further metadata.

Table 3: VV-Ghana-Small Dataset Size

Split	# Participants	# Chunks	Time
Train	79	158	0.5 h
Valid	25	50	0.1 h
Test	25	50	0.1 h
Total	129	258	0.7 h

Each participant is randomly assigned to be part of either the *training*, *validation*, or *test* set to ensure that all participants seen during validation and test are previously unseen by the model.

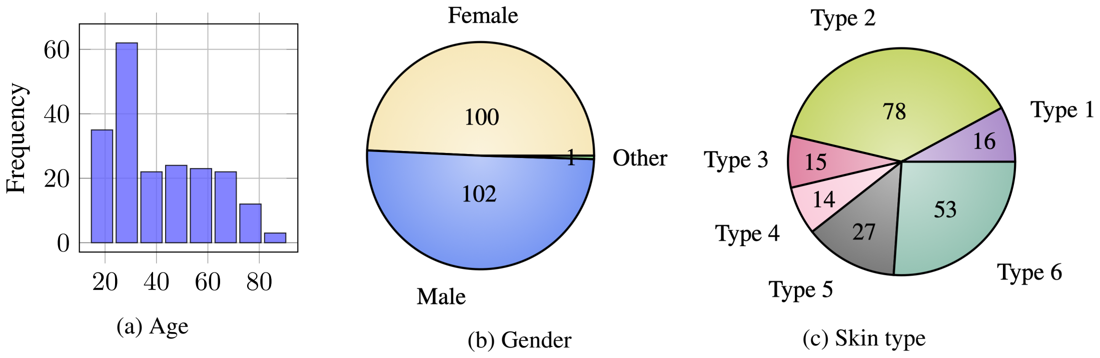
3.4 Final training dataset: PROSIT + VV-Ghana-Small

We combine the training sets of PROSIT and VV-Ghana-Small for training. As of this writing, this makes a total of 193 unique participants, 6,923 chunks or 19.9 hours of data.

Table 4: VitalLens Training Dataset Size

Source	# Participants	# Chunks	Time
PROSIT	114	6,765	19.4 h
VV-Ghana-Small	79	158	0.5 h
Total	193	6,923	19.9 h

Dataset demographics. The demographics of our training dataset are given in Figure 1. The participants are predominantly on the younger side, but as we show in Section 5, this is not an issue in practice. Genders are equally represented. Although the skin type diversity of PROSIT by itself is lacking, this combined training dataset has a diverse representation of skin types. As we show in Section 5, this helps us to reduce the skin type bias of VitalLens.



(a) Age

(b) Gender

(c) Skin type

Figure 1: Participant demographics in training dataset

Dataset vitals diversity. Distributions of the vitals in our training dataset are given in Figure 2. The participants are mostly healthy, so these vitals fall in the typical ranges. There are several participants who have an irregular heartbeat.

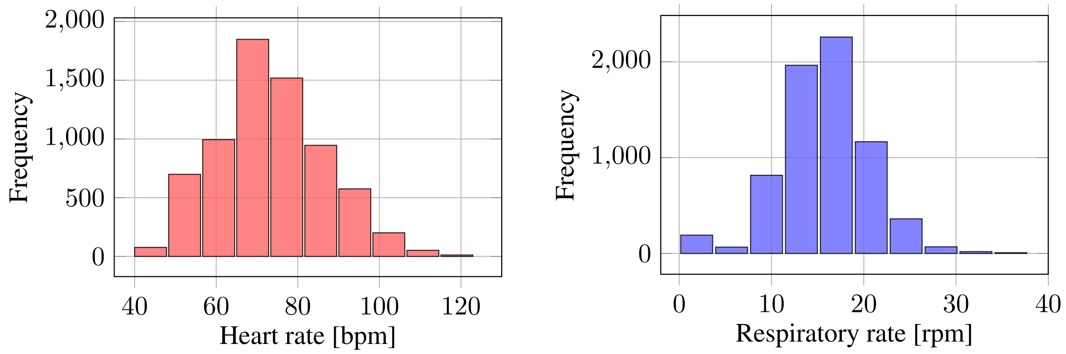


Figure 2: Distributions of chunk summary vitals in training dataset

4 Methodology

To develop and evaluate the VitalLens estimation engine model, we follow a systematic approach involving model training, validation, and testing:

Model Training. The training process optimizes the model parameters to learn the mapping between facial video signals and ground truth physiological data. For this purpose, we use a diverse training dataset of 193 unique participants introduced in Section 3.

Model Validation. While developing the model, we train many different versions of our model. During training, we continuously monitor the model's performance on the validation sets of PROSIT and VV-Ghana-Small as outlined in Section 3. Since the training dataset is disjoint from both validation sets in terms of participants, this allows us to calculate validation metrics which measure how well each model has learned to generalize in estimating vital signs from video. We used the validation metrics to select the final model to be used in VitalLens.

Model Testing. The final evaluation of the VitalLens estimation engine is conducted by measuring the generalization abilities of the final model on several datasets, all of which are participant-disjoint from both the training and validation datasets:

- The entire VV-Medium dataset, a large and diverse dataset of 289 participants introduced in Section 3. It includes participants with varying demographics and environmental conditions.
- The test set of the PROSIT dataset. This is important as PROSIT includes more camera and participant movement, which is to be expected when VitalLens is used in the wild.
- The test set of the VV-Ghana-Small dataset.

The evaluation metrics include key vital signs such as heart rate and respiratory rate. We compare the performance of VitalLens with existing methods, including G, CHROM, POS, DeepPhys, and MTTS-CAN, using metrics such as Signal-to-Noise Ratio (SNR) and Mean Absolute Error (MAE).

This methodology ensures a thorough assessment of the VitalLens estimation engine, considering diverse data sources and real-world conditions. The results and discussions in the following sections provide insights into the effectiveness and limitations of the proposed approach.

5 Results and Discussion

5.1 Vitals estimation

Table comparing G, CHROM, POS, DeepPhys, MTTS-CAN, VitalLens

5.2 Which factors impact estimation performance?

Regression analysis to determine factors impacting estimation performance

5.3 Impact of subject movement

Bar chart comparing G, CHROM, POS, DeepPhys, MTTS-CAN, VitalLens SNR or MAE estimating HR - vs subject movement on x axis. Bar chart comparing DeepPhys, MTTS-CAN, VitalLens SNR or MAE estimating RR - vs subject movement on x axis.

5.4 Impact of illumination variation

Bar chart comparing G, CHROM, POS, DeepPhys, MTTS-CAN, VitalLens SNR estimating HR - vs illumination variation on x axis.

5.5 Impact of subject skin type

Bar chart comparing G, CHROM, POS, DeepPhys, MTTS-CAN, VitalLens SNR estimating HR - vs subject skin type on x axis.

5.6 Impact of subject age

Bar chart comparing G, CHROM, POS, DeepPhys, MTTS-CAN, VitalLens SNR estimating HR - vs subject age on x axis.

6 Conclusion

Authors may wish to optionally include extra information (complete proofs, additional experiments and plots) in the appendix. All such materials should be part of the supplemental material (submitted separately) and should NOT be included in the main submission.

References

References

[1] Weixuan Chen and Daniel McDuff. Deepphys: Video-based physiological measurement using convolutional attention networks. In *ECCV*, 2018.

[2] Gerard de Haan and Vincent Jeanne. Robust pulse rate from chrominance-based rppg. *IEEE Transactions on Biomedical Engineering*, 60(10):2878–2886, 2013.

[3] Xin Liu, Josh Fromm, Shwetak Patel, and Daniel McDuff. Multi-task temporal shift attention networks for on-device contactless vitals measurement. In *NeurIPS*, 2020.

[4] Mingxing Tan and Quoc V Le. Efficientnetv2: Smaller models and faster training. In *ICML*, 2021.

[5] Pieter-Jan Toye. Vital videos: A public dataset of videos with ppg and bp ground truths. *arXiv preprint arXiv:2306.11891*, 2023.

[6] Wim Verkruysse, Lars O Svaasand, and J Stuart Nelson. Remote plethysmographic imaging using ambient light. *Optics Express*, 16(26):21434–21445, 2008.

[7] Wenjin Wang, Albertus C den Brinker, Sander Stuijk, and Gerard de Haan. Algorithmic principles of remote ppg. *IEEE Trans. Biomedical Eng.*, 64(7):1479–1491, 2017.