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 has proven to be a rich source of information about an individual's physiological state [7]. In particular, signals embedded in these videos can be harnessed to estimate vital signs such as heart rate and respiratory rate through a process known as Remote Photoplethysmography (rPPG). This capability holds immense potential for non-invasive and real-time health monitoring applications.

Various rPPG approaches have been proposed, including handcrafted algorithms and ones learned from empirical data. Handcrafted algorithms offer advantages of fast inference speed and no training data required, but usually lack in accuracy; notable approaches include the original approach G [7], and the more sophisticated CHROM [2] and POS [8]. Given enough high-quality data to learn from, learning-based approaches such as DeepPhys [1] and MTTS-CAN [4] hold the promise of greater accuracy, which they have to trade off against inference speed. There are also other learning-based approaches achieving high accuracy, which we do not cover here as they are not aiming to achieve real-time inference.

In this context, we introduce VitalLens, an innovative application designed to estimate vital signs, including heart rate and respiratory rate, in real time from selfie videos. The underlying estimation engine of VitalLens is built upon a computer vision model trained on a diverse dataset, combining video recordings with gold-standard ground truth physiological sensor data.

[INSERT FIGURE WITH SAMPLE PREDICTIONS FROM VALIDATION SET]

This report focuses on the capabilities and limitation of VitalLens. To this end, We present a comprehensive performance evaluation, comparing it with the named existing methods on diverse datasets.

The remainder of this report is organized as follows: Section 2 provides an brief overview of the model, Section 3 details the datasets used for training and evaluation, Section 4 outlines the systematic approach to model development and evaluation, Section 5 presents the results and discussions, and finally, Section 6 concludes the report with implications and future directions.

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* [5] 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 [3].

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 [6]. 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 289 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. Note that for most chunks, the missing respiratory signal was synthetically created.[1]

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. Note that for some chunks, the missing respiratory signal was synthetically created using the same procedure as for VV-Medium.

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

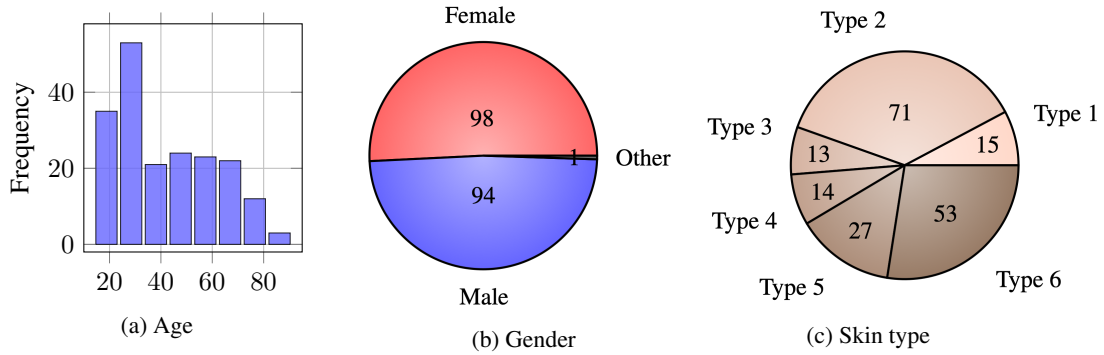
We combine the training sets of PROSIT and VV-Ghana-Small for training. As of this writing, this makes a total of 193 unique participants, 6,923 chunks or 19.9 hours of data.

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.

[1]This was done using an earlier version of our model trained on PROSIT. We then manually verified the correctness by visual inspection of both the label and video, and discarded bad labels.



(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.

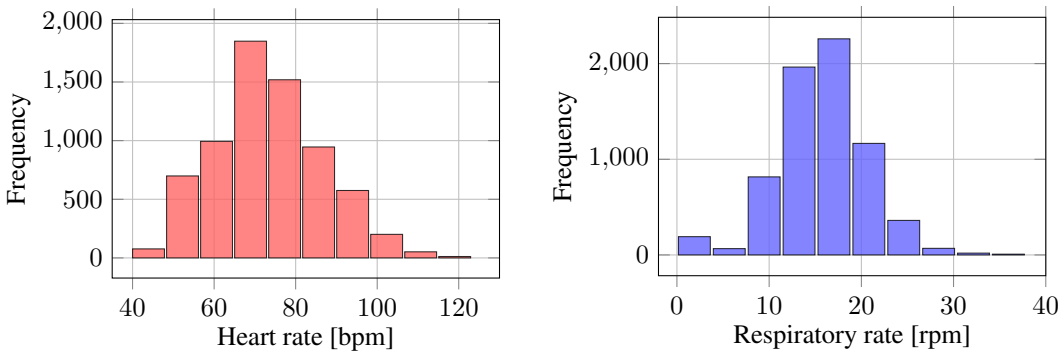


Figure 2: Distributions of chunk summary vitals in training dataset

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 benchmarking 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 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.

5 Results and Discussion

We evaluated VitalLens on three datasets - VV-Medium, the PROSIT test set, and the VV-Ghana-Small test set. None of the participants included in these datasets were previously seen by VitalLens. In addition, we also evaluated several other existing methods to allow benchmarking of VitalLens: The handcrafted methods G, CHROM, and POS, as well as the learning-based methods DeepPhys and MTTS-CAN. To allow a fair comparison, we trained the latter two on the same training data as VitalLens.

Furthermore, we conduct a regression analysis by using metadata to predict the performance of VitalLens. This helps us identify the factors affecting performance, and informs "how to use" advice offered to users of VitalLens.

Finally, we analyze the influence of the factors of interest in more detail.

5.1 Vitals estimation

Vitals estimation is performed separately for each chunk at 30 frames per second for all compared methods. We use the dataset labels and model estimations to calculate several evaluation metrics for each chunk: Mean absolute error (MAE, smaller is better), signal-to-noise ratio (SNR, larger is better), and pearson correlation coefficient (r, larger is better). We then report the mean for each of these metrics across all chunks.

Table 5: Vitals estimation results on VV-Medium

Method	Heart rate	Pulse wave		Respiratory rate[a]	Respiration wave[a]	
	MAE ↓	SNR ↑	r ↑	MAE ↓	SNR ↑	r ↑
G	10.09	−3.52	0.08	–	–	–
CHROM	5.18	0.48	0.33	–	–	–
POS	2.24	5.97	0.10	–	–	–
VitalLens	0.71	11.02	0.75	0.76	7.85	0.89

[a] Missing respiratory signals were synthetically created and manually verified.

Table 6: Vitals estimation results on PROSIT test set

Method	Heart rate	Pulse wave		Respiratory rate	Respiration wave	
	MAE ↓	SNR ↑	r ↑	MAE ↓	SNR ↑	r ↑
G	27.71	−15.01	0.22	–	–	–
CHROM	24.06	−11.67	0.16	–	–	–
POS	17.79	−11.94	0.28	–	–	–
VitalLens	4.62	2.83	0.67	3.76	−1.23	0.47

Table 7: Vitals estimation results on VV-Ghana-Small test set

Method	Heart rate	Pulse wave		Respiratory rate[a]	Respiration wave[a]	
	MAE ↓	SNR ↑	r ↑	MAE ↓	SNR ↑	r ↑
G	24.50	−12.48	0.13	–	–	–
CHROM	18.90	−8.03	0.23	–	–	–
POS	10.63	−6.20	0.15	–	–	–
VitalLens	1.71	9.05	0.79	2.43	4.40	0.81

[a] Missing respiratory signals were synthetically created and manually verified.

5.2 Which factors impact heart rate estimation performance?

We consider a number of factors that may impact heart rate estimation performance:

- `age`: The age of the participant
- `camera_stationary`: Indicates whether the camera is stationary.
- `gender_male`: Dummy variable indicating whether the participant is male.
- `illuminance_var`: Measures how much the illuminance of the participant's faces varies throughout a chunk, in interval [0,1].
- `movement`: Measures how much the participant moved throughout the chunk, in interval [0,1].
- `skin_type_x`: Dummy variable indicating whether the participant has skin type x. Base case is skin type 1.

To investigate which of these factors have a significant impact, we conduct OLS regressions using them to predict the heart rate estimation MAE. We perform separate regressions using the chunks of the VV-Medium dataset and the PROSIT test set - the results are reported in Tables 8 and 9.

Table 8: Regression analysis of factors affecting heart rate estimation on VV-Medium

Dep. Variable:	HR MAE	**R-squared:**	0.075
Model:	OLS	**Adj. R-squared:**	0.068
Method:	Least Squares	**F-statistic:**	9.914
No. Observations:	1108	**Prob (F-statistic):**	1.00e-14
Df Residuals:	1098	**AIC:**	-419.0
Df Model:	9	**BIC:**	-368.9

	coef	**std err**	**t**	**P$>$ |t|**	**[0.025**	**0.975]**
intercept	0.5320	0.035	15.257	0.000	0.464	0.600
age	-0.0014	0.000	-4.658	0.000	-0.002	-0.001
illuminance_var	0.3539	0.136	2.594	0.010	0.086	0.622
movement	0.1967	0.106	1.855	0.064	-0.011	0.405
gender_male	0.0098	0.012	0.803	0.422	-0.014	0.034
skin_type_2	0.0107	0.034	0.315	0.753	-0.056	0.078
skin_type_3	-0.0511	0.037	-1.390	0.165	-0.123	0.021
skin_type_4	-0.0565	0.039	-1.458	0.145	-0.132	0.020
skin_type_5	0.0164	0.038	0.433	0.665	-0.058	0.091
skin_type_6	0.1486	0.041	3.594	0.000	0.067	0.230

The regression for VV-Medium shows a significant result. Only `age`, `illuminance_var`, and `skin_type_6` are shown to have significant effects at the 5% level:

- `age` has a weak negative effect, meaning that estimations were slightly more accurate for older participants,
- `illuminance_var` had a moderate positive effect, meaning that higher variance in facial illumination of participants led to less accurate estimations, and
- `skin_type_6` had a moderate positive effect, meaning that the error of heart rate estimation for participants with skin type 6 was on average ca. 0.15 bpm higher, taking into account the other factors.

For VV-Medium, the factors are collectively able to explain a modest 7.5% of the variance of the absolute errors in HR estimation.

Looking at the regression for the PROSIT test set, we find that it has much more explanatory power. Here, the regression itself and all considered factors appear to have significant effects at the 5% level:

Table 9: Regression analysis of factors affecting heart rate estimation on PROSIT test set

Dep. Variable:	HR MAE	**R-squared:**	0.350
Model:	OLS	**Adj. R-squared:**	0.346
Method:	Least Squares	**F-statistic:**	93.64
No. Observations:	1403	**Prob (F-statistic):**	1.58e-124
Df Residuals:	1394	**AIC:**	-231.2
Df Model:	8	**BIC:**	-184.0

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
intercept	0.5039	0.028	17.823	0.000	0.448	0.559
age	-0.0041	0.000	-9.169	0.000	-0.005	-0.003
illuminance_var	0.6905	0.075	9.201	0.000	0.543	0.838
movement	0.3509	0.039	8.989	0.000	0.274	0.428
camera_stationary	-0.1748	0.017	-10.056	0.000	-0.209	-0.141
gender_male	0.1722	0.014	11.998	0.000	0.144	0.200
skin_type_2	0.1268	0.016	8.013	0.000	0.096	0.158
skin_type_3	-0.1226	0.058	-2.124	0.034	-0.236	-0.009
skin_type_4	0.3286	0.039	8.391	0.000	0.252	0.405

- `age` has a weak negative effect, same direction but stronger than the result on VV-Medium. This means that again, estimations were slightly more accurate for older participants.
- `camera_stationary` had a negative effect, meaning that estimations were better when the camera was stationary.
- `gender_male` has a moderately positive effect, meaning that estimations were worse for male participants.
- `illuminance_var` had a strong positive effect, meaning that estimations were worse with higher variance in facial illumination of participants.
- `movement` had a positive effect, meaning that estimations were worse when participants moved more.
- `skin_type_x`: Only skin types 1, 2, 3, and 4 were present in the PROSIT test set and the result is mixed, so there is no clear conclusion.

For PROSIT, the factors are collectively able to explain a respectable 35% of the variance of the absolute errors in HR estimation.

These results confirm that both participant movement and variation in the illumination of the participant's face have a negative impact on estimation performance. However, we note that these effects are contained to less than 1 bpm from the best to worst values of the respective factors. Secondly, it is interesting to note that at least for these datasets and the our estimation model, the effect of illumination variation seems to have a greater impact than participant movement. Additionally, we observe that estimation performance appears to be better for female participants. On the one hand, this may be influenced by bearded male participants, which we don't control for – however, on the other hand, we also don't control for use of make-up by female participants.

Based on these results, we advise users of VitalLens that for best performance they should (i) hold still and place the device on a surface, and (ii) make sure the face is evenly illuminated.

5.3 Which factors impact respiratory rate estimation performance?

For our analysis of the impact on respiratory rate estimation, we consider the factors `age`, `camera_stationary`, `gender`, `illuminance_var`, and `movement`.

5.4 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.

Table 10: Regression analysis of factors affecting respiratory rate estimation on PROSIT test set

Dep. Variable:	resp_mae_vl	**R-squared:**	0.126
Model:	OLS	**Adj. R-squared:**	0.122
Method:	Least Squares	**F-statistic:**	40.11
No. Observations:	1403	**Prob (F-statistic):**	1.28e-38
Df Residuals:	1397	**AIC:**	396.3
Df Model:	5	**BIC:**	427.8

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
intercept	0.7648	0.031	24.842	0.000	0.704	0.825
age	-0.0025	0.001	-4.849	0.000	-0.004	-0.001
illuminance_var	0.5166	0.091	5.682	0.000	0.338	0.695
movement	0.5376	0.048	11.210	0.000	0.444	0.632
camera_stationary	-0.0351	0.020	-1.731	0.084	-0.075	0.005
gender_male	-0.0076	0.016	-0.481	0.630	-0.039	0.023

5.5 Impact of illumination variation

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

5.6 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.

5.7 Impact of subject age

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

5.8 Impact of vitals

Investigate whether predictions are worse for lower / higher HR

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] Thomas B Fitzpatrick. Soleil et peau. *Journal de Médecine Esthétique*, 2, 1975.

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

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

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

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

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