
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. VitalLens uses a computer vision model trained on a diverse dataset of video and physiological sensor data. We benchmark performance on several diverse datasets, including VV-Medium, which consists of 289 unique participants. VitalLens outperforms several existing methods including POS and MTTs-CAN on all datasets while maintaining a fast inference speed. On VV-Medium, VitalLens achieves mean absolute errors of 0.71 bpm for heart rate estimation, and 0.76 bpm for respiratory rate estimation.

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 [8]. In particular, signals embedded in these videos can be harnessed to estimate vital signs 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 [8], and the more sophisticated CHROM [2] and POS [9]. 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 is traded off against inference speed. There are also other learning-based approaches, 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. This report focuses on the capabilities and limitation of VitalLens. Our key contributions are:

- **Real-time rPPG application.** We introduce VitalLens – the first widely distributed rPPG application for real time estimation of vitals such as of heart rate and respiratory rate.¹
- **Comprehensive evaluation on a large dataset.** We evaluate the performance of VitalLens on multiple datasets, including VV-Medium [7] with 289 participants. VitalLens outperforms G, CHROM, POS, DeepPhys, and MTTs-CAN in terms of estimation accuracy. On VV-Medium, VitalLens achieves mean absolute errors of 0.71 bpm for heart rate estimation, and 0.76 bpm for respiratory rate estimation, while taking 18 ms to run inference.
- **Analysis of factors impacting performance.** We systematically investigate factors impacting estimation performance. We find that *participant movement* and *variation of participant illuminance* are the main factors impacting performance. Compared to approaches like POS, VitalLens is less impacted by darker skin types when estimating heart rates.
- **Maintaining end-user privacy.** VitalLens runs inference locally, meaning that video and vitals never leave the device.

¹Click to get VitalLens for iOS.

2 Architecture

VitalLens takes a series of video frames as inputs, and uses them to estimate the pulse waveform and the respiration waveform. The corresponding heart rate (HR) and respiratory rate (RR) estimates are derived from these waveforms using Fast Fourier Transform (FFT). Please refer to Figure 1 for a broad overview of how VitalLens operates, along with the evaluation method used in Section 5.

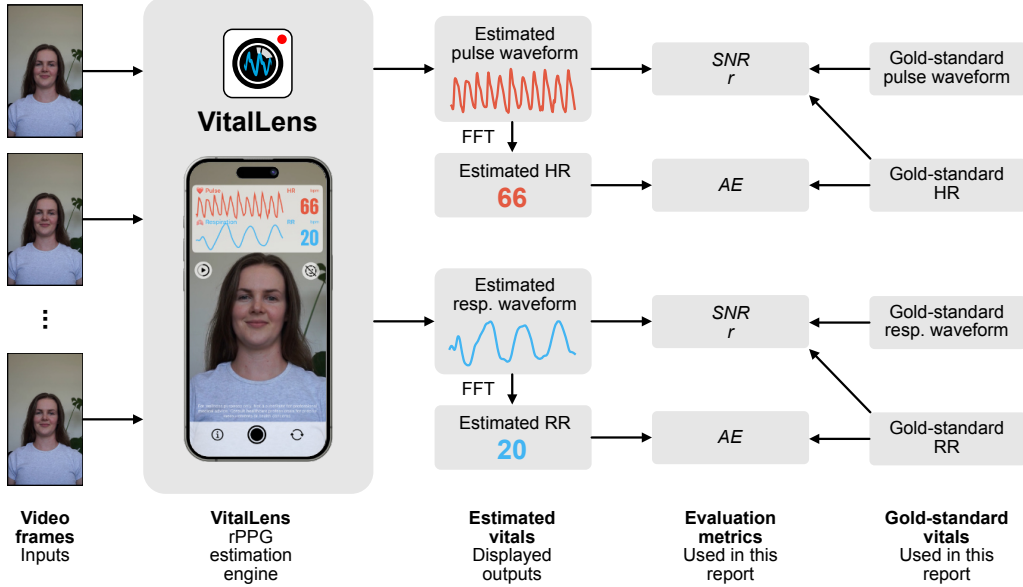


Figure 1: VitalLens estimates vital sign waveforms from video frames. The app displays the video feed with an overlay of the estimated pulse and respiration waveforms as well as the derived heart rate and respiration rate. We report evaluation metrics for all displayed quantities by comparing the estimations with gold-standard labels.

The estimation engine built into VitalLens (henceforth “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* [6] 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 Africa* 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 Africa* 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 illuminance 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 [7]. 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.²

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 Africa

Vital Videos Africa 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-Africa-Small”), which consists of 129 participants. Both camera and participants are stationary in this subset.

²This 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.

Pre-processing. We pre-process VV-Africa-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-Africa-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-Africa-Small

We combine the training sets of PROSIT and VV-Africa-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-Africa-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 2. 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.

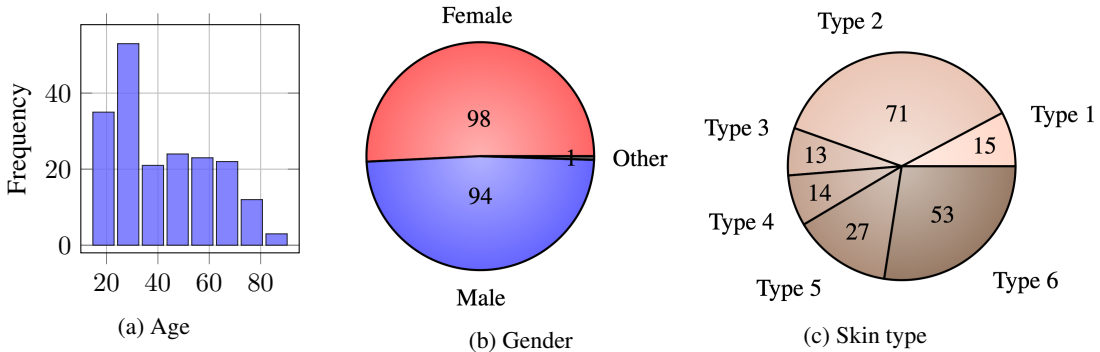


Figure 2: Participant demographics in training dataset

Dataset vitals diversity. Distributions of the vitals in our training dataset are given in Figure 3. The participants are mostly healthy, so these vitals fall in the typical ranges. There are several participants who have an irregular heartbeat.

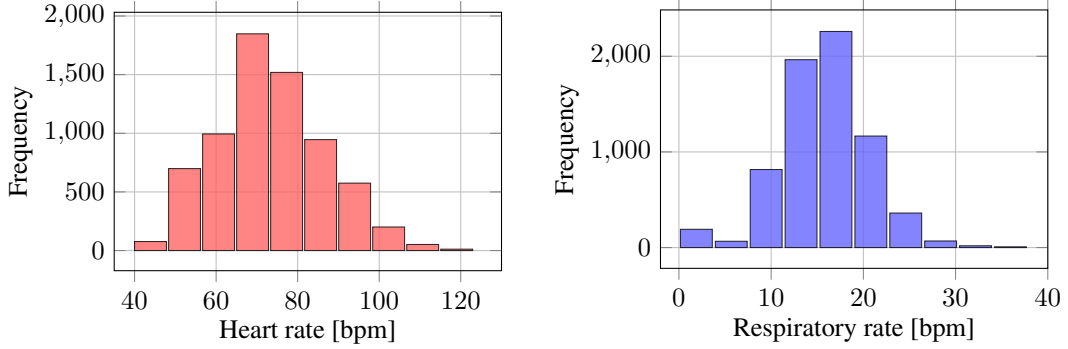


Figure 3: Distributions of chunk summary vitals in training dataset

4 Methodology

To develop and evaluate VitalLens, 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-Africa-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 VitalLens 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-Africa-Small dataset.

Inference and Privacy. When end-users use VitalLens on their devices, inference runs locally. This means that the video and vitals of end-users never leave their devices, which is important to maintain privacy. Another benefit is that an internet connection is not required to run VitalLens.

5 Results and Discussion

We evaluated VitalLens on three datasets - VV-Medium, the PROSIT test set, and the VV-Africa-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. As shown in Figure 1, we use the dataset labels and model estimations to calculate several evaluation metrics for each chunk:

- **Absolute error** (AE, smaller is better): Used to measure the absolute difference between the estimated rate and the true rate.
- **Signal-to-noise ratio** (SNR, larger is better): Measures the ratio of the true rate signal to the noise present in the estimated waveform in decibels.
- **Pearson correlation coefficient** (r , larger is better): Measures the the linear correlation between the estimated waveform and the true waveform.

We then report the mean for each of these metrics across all chunks in the given dataset. As is common, we denote the mean AE as MAE as is common, but continue to use the abbreviations SNR and r when reporting the mean SNR and mean r .

Table 5: Vitals estimation results on VV-Medium

Method	Heart rate	Pulse wave		Respiratory rate ^a	Respiration wave ^a		Inference time in ms ^b
	MAE ↓	SNR ↑	r ↑	MAE ↓	SNR ↑	r ↑	
G	10.09	−3.52	0.08	—	—	—	3.4
CHROM	5.18	0.48	0.33	—	—	—	4.2
POS	2.24	5.97	0.10	—	—	—	3.6
DeepPhys	9.62	−2.07	0.52	2.06	4.22	0.73	9.8
MTTS-CAN	1.31	6.30	0.71	5.36	−4.25	0.43	22.1
VitalLens	0.71	11.02	0.75	0.76	7.85	0.89	18.0

^a Missing respiratory signals were synthetically created and manually verified.

^b Time to run inference for a single video frame.

Results for VV-Medium are reported in Table 5. The heart rate and respiratory rate estimated by VitalLens across all chunks was off by 0.71 bpm and 0.76 bpm on average from the gold-standard labels, while taking 18 ms for inference.

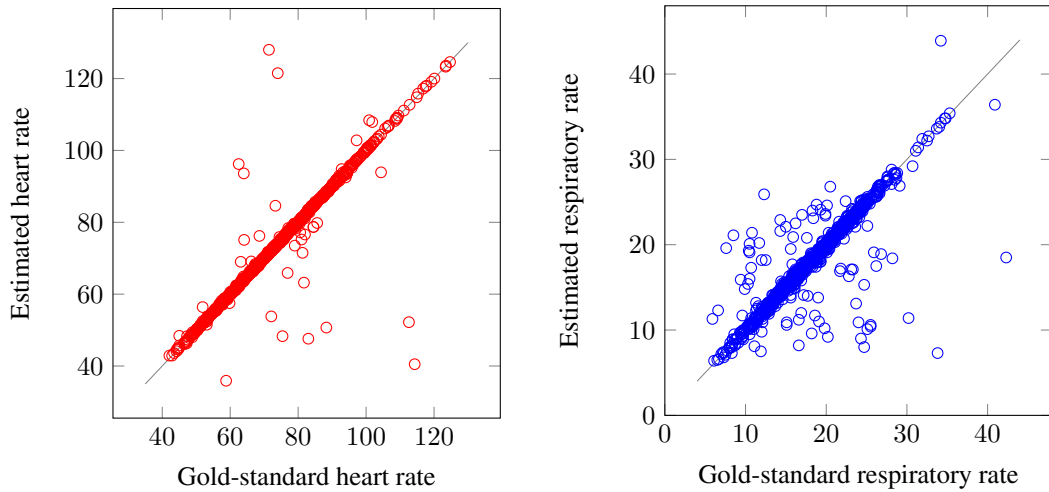


Figure 4: VitalLens estimated vitals vs. gold-standard true vitals on VV-Medium. Besides a few outliers, estimations closely match the true vitals.

VitalLens beats all other listed methods for all considered evaluation metrics, while maintaining a competitive inference speed that is sufficient to run in real time at 30 fps.³ The next best predictions for heart rate and respiratory rate also came from learning-based models MTTs-CAN and DeepPhys, respectively. Note that this dataset has a large variety of participants and their demographics, but low amounts of participant and camera movement.

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	—	—	—
DeepPhys	29.53	−12.41	0.11	4.95	−4.09	0.40
MTTS-CAN	13.90	−6.73	0.33	5.95	−6.48	0.23
VitalLens	4.62	2.83	0.67	3.76	−1.23	0.47

The PROSIT test set has a lower variety in participants and their demographics, but it includes both participant and camera movement. Results for the PROSIT test set are reported in Table 6 – again, VitalLens beats all listed methods for all considered evaluation metrics. However, we observe that here, the absolute errors achieved by VitalLens for heart rate and respiratory rate are 4.62 bpm and 3.76 bpm, respectively – markedly higher than for VV-Medium. This is a first hint that factors like participant and camera movement are the main drivers of lower estimation accuracies.

Table 7: Vitals estimation results on VV-Africa-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	—	—	—
DeepPhys	26.79	−8.54	0.29	2.57	3.56	0.75
MTTS-CAN	5.79	−2.91	0.51	4.10	−2.72	0.34
VitalLens	1.71	9.05	0.79	2.43	4.40	0.81

^a Missing respiratory signals were synthetically created and manually verified.

Finally, Table 7 lists the results for the VV-Africa-Small test set. The results paint a similar picture as the two previous datasets, with VitalLens beating out the other methods. In comparison to VV-Medium, we observe that the errors are generally higher - in this case, this may be caused by the demographic differences between the participants and more challenging lighting conditions. We further investigate this in the following sections.

5.2 Which factors impact estimation performance?

We consider a number of factors that may impact estimation performance of pulse and respiration.

Demographic factors:

age: The age of the participant.

gender_male: Dummy variable indicating whether the participant is male.

hr / rr: The heart rate and respiratory rate of the participant.

skin_type_x: Dummy variable indicating whether the participant has skin type x. Base case is skin type 1.

³The inference times are for a single video frame, excluding face detection but including all other computation, such as resizing. We measure these times using our own implementations as an average across 1,000 iterations on a workstation with Nvidia GeForce RTX 3080 and AMD Ryzen 7 3700X.

Behavioral/environmental factors:

`camera_stationary`: Indicates whether the camera is stationary.

`illuminance_var`: Measures how much the illuminance of the participant's faces varies throughout a chunk, in interval [0,1]. Derived from the luma component of the pixels within the participant's face. Partially caused by movement ($r = 0.2$ for PROSIT).

`movement`: Measures how much the participant moved throughout the chunk, in interval [0,1]. Derived from the movement of facial landmarks.

5.2.1 Pulse wave

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

Table 8: Regression analysis of factors affecting pulse wave estimation on VV-Medium

Dep. Variable:	Pulse wave SNR	R-squared:	0.264
Model:	OLS	Adj. R-squared:	0.257
Method:	Least Squares	F-statistic:	39.27
No. Observations:	1108	Prob (F-statistic):	2.40e-66
Df Residuals:	1097	AIC:	7118.
Df Model:	10	BIC:	7174.

	coef	std err	t	P> t	[0.025	0.975]
intercept	9.7449	1.575	6.189	0.000	6.655	12.834
age	0.1135	0.010	11.919	0.000	0.095	0.132
illuminance_var	-22.3650	4.096	-5.460	0.000	-30.403	-14.327
movement	-5.3373	3.179	-1.679	0.093	-11.576	0.901
hr	-0.0283	0.014	-2.046	0.041	-0.055	-0.001
gender_male	0.2530	0.377	0.672	0.502	-0.486	0.992
skin_type_2	-0.1885	1.024	-0.184	0.854	-2.199	1.822
skin_type_3	0.4203	1.104	0.381	0.703	-1.745	2.586
skin_type_4	1.4931	1.161	1.286	0.199	-0.785	3.772
skin_type_5	-2.6020	1.140	-2.282	0.023	-4.839	-0.365
skin_type_6	-7.9805	1.241	-6.429	0.000	-10.416	-5.545

For our regression analysis on the VV-Medium, we use all available factors. The regression equation is significant and explains 26.4% of the variance of the pulse wave estimation SNR. `age`, `hr`, `illuminance_var`, and `skin_type_x` are shown to have significant effects at the 5% level:

- `age` has a weak positive effect, meaning that estimations were slightly more accurate for older participants,
- `hr` has a weak negative effect, meaning that estimations were slightly less accurate for participants with higher heart rates,
- `illuminance_var` had a strong negative effect, meaning that higher variance in facial illuminance of participants led to less accurate estimations, and
- `skin_type_5` and especially `skin_type_6` had negative effects, meaning that pulse wave estimation for participants with skin types 5 and especially 6 was less accurate.

It is worth noting that `movement` did not have a significant effect, potentially because participants generally do not move much for this dataset.

Since there are not many unique participants in the PROSIT test set, we only consider the behavioral and environmental factors for the next regression. Looking at the results, the regression equation itself as well as `illuminance_var` and `movement` have significant effects at the 5% level:

Table 9: Regression analysis of factors affecting pulse wave estimation on PROSIT test set

Dep. Variable:	Pulse wave SNR	R-squared:	0.103
Model:	OLS	Adj. R-squared:	0.102
Method:	Least Squares	F-statistic:	53.82
No. Observations:	1403	Prob (F-statistic):	6.40e-33
Df Residuals:	1399	AIC:	1.040e+04
Df Model:	3	BIC:	1.042e+04

	coef	std err	t	P> t	[0.025	0.975]
intercept	7.2056	0.669	10.775	0.000	5.894	8.517
illuminance_var	-26.4655	3.216	-8.230	0.000	-32.774	-20.157
movement	-14.3836	1.687	-8.525	0.000	-17.693	-11.074
camera_stationary	0.1855	0.597	0.311	0.756	-0.985	1.356

- `illuminance_var` has a strong negative effect, meaning that estimations became less accurate with higher variance in facial illuminance of participants.
- `movement` has a negative effect, meaning that estimations were less accurate when participants moved more.

These behavioral factors are collectively able to explain 10.3% of the variance of the pulse wave estimation SNR. It is worth noting that `camera_stationary` did not have a significant effect, which indicates that our rPPG model has some degree of robustness against camera movement.

Overall, the results confirm that the behavioral factors of participant movement and variation in the illuminance of the participant’s face dominate the results with a significant negative impact on estimation performance. It is interesting to note that at least for these datasets, the effect of illuminance variation seems to have a greater impact than participant movement. Our model performance appears to be slightly impacted by the heart rate itself. We also confirm previous findings that rPPG models can struggle with darker skin types, but cannot confirm findings that estimation is more accurate for females [5]. Additionally, we observe that estimation appears to be more accurate for older participants.

5.2.2 Respiration wave

For our analysis of the impact on respiration wave estimation, we conduct another two OLS regressions. The results are reported in Table 10 and 11.

For our regression analysis on VV-Medium, we use all available factors. The regression equation is significant but explains only 7.1% of the variance of the respiratory wave estimation SNR. `age`, `movement`, and `rr` are shown to have significant effects at the 5% level:

- `age` has a weak positive effect, meaning that estimations were slightly more accurate for older participants,
- `movement` had a strong negative effect, meaning that higher participant movement led to less accurate estimations, and
- `rr` has a weak negative effect, meaning that estimations were slightly less accurate for participants with higher respiratory rates.

For the regression on the PROSIT test set, we again focus on the behavioral and environmental factors. The regression itself and all considered factors have significant effects at the 5% level:

- `camera_stationary` has a weak negative effect, meaning that estimations were slightly less accurate when camera was stationary,
- `illuminance_var` has a strong negative effect, meaning that estimations were less accurate when participant illuminance varied more, and
- `movement` had a strong negative effect, meaning that higher participant movement led to less accurate estimations.

Table 10: Regression analysis of factors affecting respiratory rate estimation on VV-Medium

Dep. Variable:	Resp wave SNR	R-squared:	0.071
Model:	OLS	Adj. R-squared:	0.062
Method:	Least Squares	F-statistic:	8.340
No. Observations:	1108	Prob (F-statistic):	3.71e-13
Df Residuals:	1097	AIC:	7352.
Df Model:	10	BIC:	7407.

	coef	std err	t	P> t	[0.025	0.975]
intercept	8.6117	1.349	6.385	0.000	5.965	11.258
age	0.0674	0.010	6.559	0.000	0.047	0.088
illuminance_var	-4.6049	4.564	-1.009	0.313	-13.560	4.350
movement	-13.8675	3.536	-3.922	0.000	-20.805	-6.930
rr	-0.1615	0.040	-4.056	0.000	-0.240	-0.083
gender_male	0.7883	0.407	1.939	0.053	-0.009	1.586
skin_type_2	-0.8223	1.139	-0.722	0.470	-3.057	1.413
skin_type_3	-0.5369	1.230	-0.437	0.663	-2.950	1.876
skin_type_4	0.1492	1.296	0.115	0.908	-2.395	2.693
skin_type_5	1.1431	1.269	0.901	0.368	-1.346	3.632
skin_type_6	-0.3998	1.383	-0.289	0.773	-3.114	2.314

Table 11: Regression analysis of factors affecting respiratory rate estimation on PROSIT test set

Dep. Variable:	Resp wave SNR	R-squared:	0.068			
Model:	OLS	Adj. R-squared:	0.066			
Method:	Least Squares	F-statistic:	33.96			
No. Observations:	1403	Prob (F-statistic):	3.46e-21			
Df Residuals:	1399	AIC:	9735.			
Df Model:	3	BIC:	9756.			
	coef	std err	t	P> t 	[0.025	0.975]
intercept	2.4654	0.527	4.676	0.000	1.431	3.500
illuminance_var	-11.0138	2.535	-4.344	0.000	-15.987	-6.040
movement	-11.9294	1.330	-8.968	0.000	-14.539	-9.320
camera_stationary	-1.3251	0.471	-2.816	0.005	-2.248	-0.402

The factors are collectively able to explain 6.8% of the variance of the respiratory wave estimation SNR. As expected, participant movement degrades the estimation quality for the respiratory rate. It is surprising that variation in the illuminance of the participant’s face impacts the estimation to a similar degree. Also, the finding that estimations were more accurate when the camera was not stationary is counterintuitive. One possible explanation is that VitalLens learns to understand that for some participants, repetitive movement of the (handheld) camera actually reveals information about respiration.

Based on the results for both pulse and respiration, we advise users of VitalLens that for best performance they should (i) hold still, and (ii) make sure the face is evenly illuminated.

5.2.3 Impact of participant movement

According to the regression analysis, increasing amounts of participant movement lead to worse estimation performances for both heart rate and respiratory rate. We explore this effect in Figure 5 by looking the average SNR for different buckets of participant movement on the PROSIT test set. It is evident that SNR values indeed decrease with increased movement. There appears to be a large drop-off especially going from chunks that have no movement to those that have a few movements.

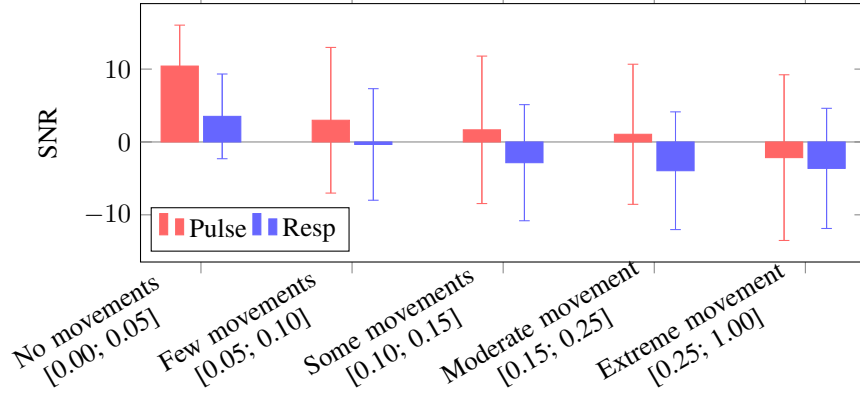


Figure 5: Signal-to-noise ratio for pulse and respiration on the PROSIT test set, grouped by different levels of participant movement.

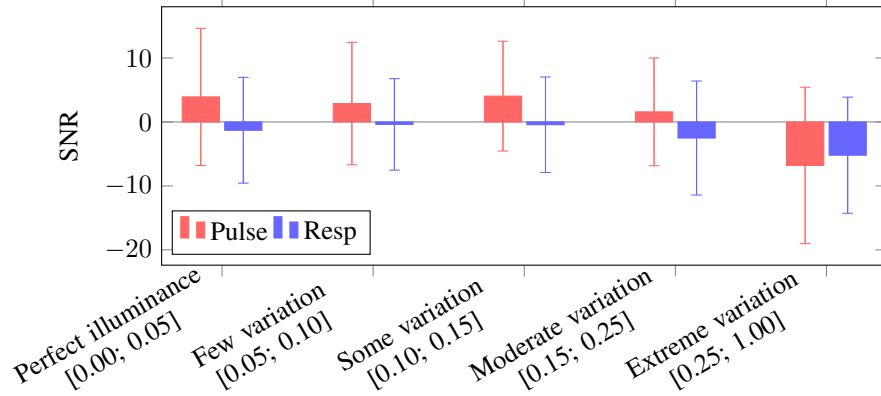


Figure 6: Signal-to-noise ratio for pulse and respiration on the PROSIT test set, grouped by different levels of participant illuminance variation.

5.2.4 Impact of participant illuminance variation

The regression analysis showed that increasing amounts of variation in participant illuminance also led to worse estimation performances for both heart rate and respiratory rate. We explore this effect in Figure 6 by looking at the average SNR for different buckets of variation in participant illuminance on the PROSIT test set. This figure reveals that especially when going from chunks with moderate illuminance variation to those with extreme illuminance variation, we see a large drop-off in estimation accuracy.

5.2.5 Impact of participant age

Since the regression results indicated that VitalLens is slightly more accurate overall for older participants, we explore this effect in more detail as well. Figure 7 indeed shows that especially for participants under 20 years of age, the SNR for both pulse and respiration is tends to be lower.

5.2.6 Impact of participant skin type

According to the regression analysis on VV-Medium, the pulse wave estimation had a lower accuracy especially for participants with skin type 6. Here, we seek to further explore this effect. We start by grouping the chunks in VV-Medium by skin type and then plot the average estimation SNR for both pulse wave and respiration wave estimation in Figure 8. Indeed, we observe that consistent

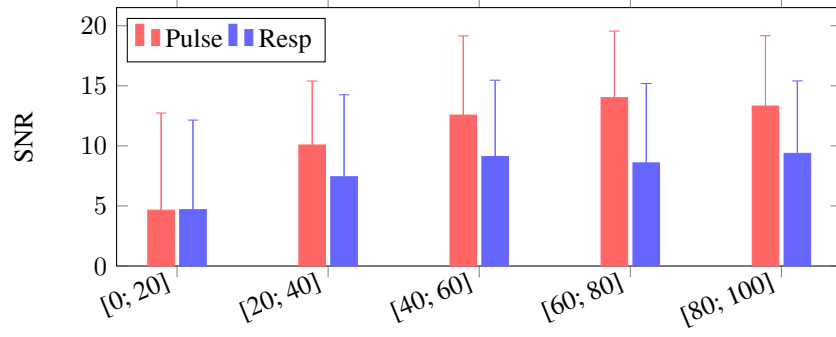


Figure 7: Signal-to-noise ratio for pulse and respiration on VV-Medium, grouped by different participant ages.

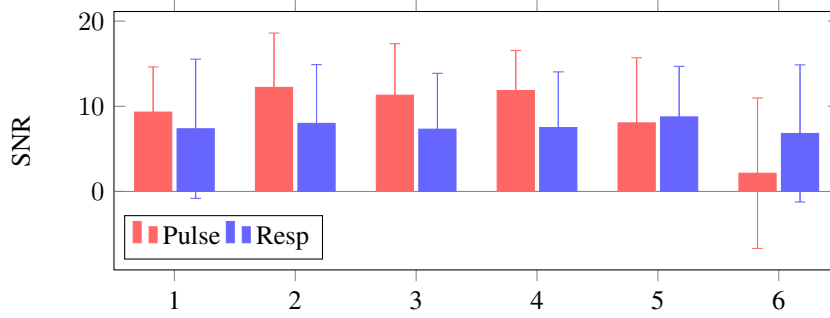


Figure 8: Signal-to-noise ratio for pulse and respiration on VV-Medium, grouped by different participant skin types.

with our previous regression analysis, there is a drop-off in the SNR between skin type 5 and 6 for pulse, while there is no obvious effect visible for respiration.

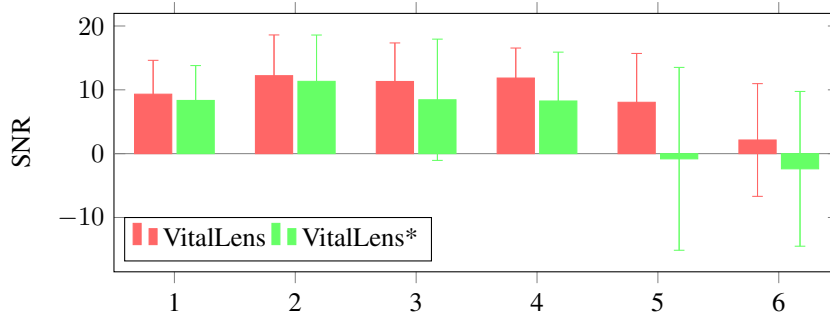


Figure 9: Comparing signal-to-noise ratio between VitalLens and VitalLens* on VV-Medium, grouped by participant skin types. VitalLens* was solely trained on PROSIT, while VitalLens was trained on both PROSIT and VV-Africa-Small.

Next, we want to investigate to what degree the inclusion of VV-Africa-Small in our training data lessened this skin type effect. For this purpose, we compare VitalLens with *VitalLens**, a version of VitalLens that is solely trained on PROSIT. The training set of PROSIT by itself only includes 2 individuals of skin type 6, a number that is massively increased in the combined training data. Both VitalLens and VitalLens* are then evaluated on VV-Medium. Figure 9 compares the results, again grouped by skin type. It is evident that the addition of VV-Africa-Small led to a sizeable improvement mainly for skin types 5 and 6, but also skin types 3 and 4.

Finally, we compare the impact of skin types on pulse wave estimation between our learned model VitalLens and the best-performing handcrafted model POS. As visible in Figure 10, the impact of

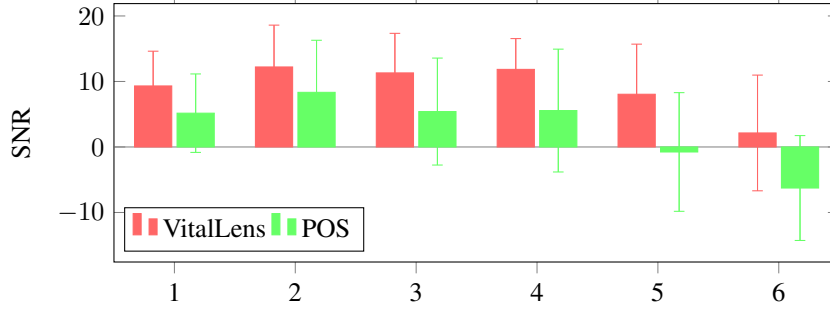


Figure 10: Comparing signal-to-noise ratio between VitalLens and POS on VV-Medium, grouped by participant skin types.

skin types in POS is even worse than on VitalLens*, with an even larger drop-off in SNR between skin types 5 and 6.

We can conclude from this analysis that issues arising from the lower pulse signal strength in rPPG for darker skin types can at least partially be overcome by learning-based rPPG models which use training data that actually includes a significant amount of individuals with those skin types.

5.2.7 Impact of participant vitals

Finally, we investigate the impact of the vitals themselves on estimation performance. Although we took steps to prevent this, the distribution of vitals in the training set may have biased VitalLens towards predicting the average values found in the training set. The regressions found that estimations generally became slightly less accurate for individuals with higher heart rates and respiratory rates. Figures 11 and 12 show the SNR for pulse and respiration estimation, grouped by different heart rate and respiratory rates respectively.

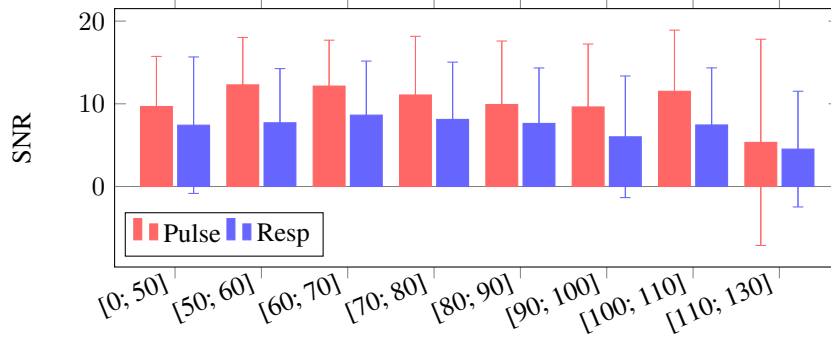


Figure 11: Signal-to-noise ratio for pulse and respiration on VV-Medium, grouped by different participant heart rate values.

The figures show that generally, VitalLens does not appear to be biased by the distributions of heart rate and respiratory rate in the training data. We do however note that pulse estimation became less accurate for the chunks with the highest heart rates, and respiration estimation became less accurate for the chunks with the highest respiration rates. This consistent with the findings of the regression analysis.

6 Conclusion

VitalLens emerges as a novel solution in the domain of rPPG, showcasing its potential for accurate and real-time estimation of vital signs from selfie videos. The extensive evaluation establishes its superiority over established methods, such as G, CHROM, POS, DeepPhys, and MTTs-CAN. Notably, VitalLens outperforms these methods while maintaining fast inference times. The thorough

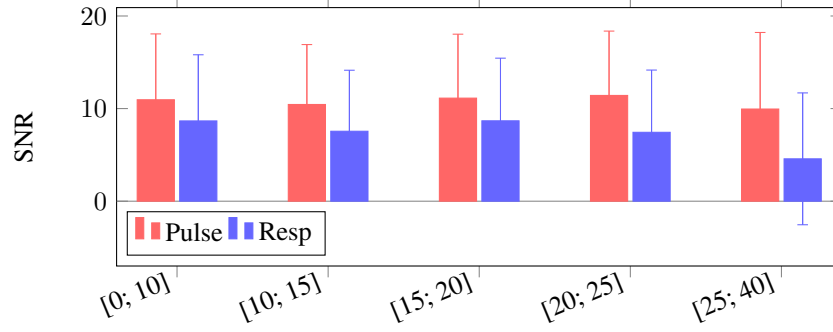


Figure 12: Signal-to-noise ratio for pulse and respiration on VV-Medium, grouped by different participant respiratory rate values.

analysis of factors impacting performance reveals the significance of participant movement and illuminance variation, providing valuable insights for optimal usage. Additionally, the inclusion of diverse training data to address skin type bias, contributes to the model’s robustness.

Acknowledgments and Disclosure of Funding

We are thankful to Pieter-Jan Toye and Andrew Goodwin for the helpful suggestions. This work was funded by Rouast Labs Pty Ltd.

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] Ewa Magdalena Nowara, Daniel McDuff, and Ashok Veeraraghavan. A meta-analysis of the impact of skin type and gender on non-contact photoplethysmography measurements. In *CVPR*, 2020.
- [6] Mingxing Tan and Quoc V Le. Efficientnetv2: Smaller models and faster training. In *ICML*, 2021.
- [7] Pieter-Jan Toye. Vital videos: A public dataset of videos with ppg and bp ground truths. *arXiv preprint arXiv:2306.11891*, 2023.
- [8] Wim Verkruijsse, Lars O Svaasand, and J Stuart Nelson. Remote plethysmographic imaging using ambient light. *Optics Express*, 16(26):21434–21445, 2008.
- [9] 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.