A Meta-Analysis of the Impact of Skin Type and Gender on Non-contact Photoplethysmography Measurements

Ewa M. Nowara*, Daniel McDuff†, Ashok Veeraraghavan*
*Rice University, Houston, TX
†Microsoft Research, Redmond, WA
{emn3,vashok}@rice.edu, damcduff@microsoft.com

Abstract

It is well established that many datasets used for computer vision tasks are not representative and may be biased. The result of this is that evaluation metrics may not reflect real-world performance and might expose some groups (often minorities) to greater risks than others. Imaging photoplethysmography is a set of techniques that enables non-contact measurement of vital signs using imaging devices. While these methods hold great promise for low-cost and scalable physiological monitoring, it is important that performance is characterized accurately over diverse populations. We perform a meta-analysis across three datasets, including 73 people and over 400 videos featuring a broad range of skin types to study how skin types and gender affect the measurements. While heart rate measurement can be performed on all skin types under certain conditions, we find that average performance drops significantly for the darkest skin type. We also observe a slight drop in the performance for females. We compare supervised and unsupervised learning algorithms and find that skin type does not impact all methods equally. The imaging photoplethysmography community should devote greater efforts to addressing these disparities and collecting representative datasets.

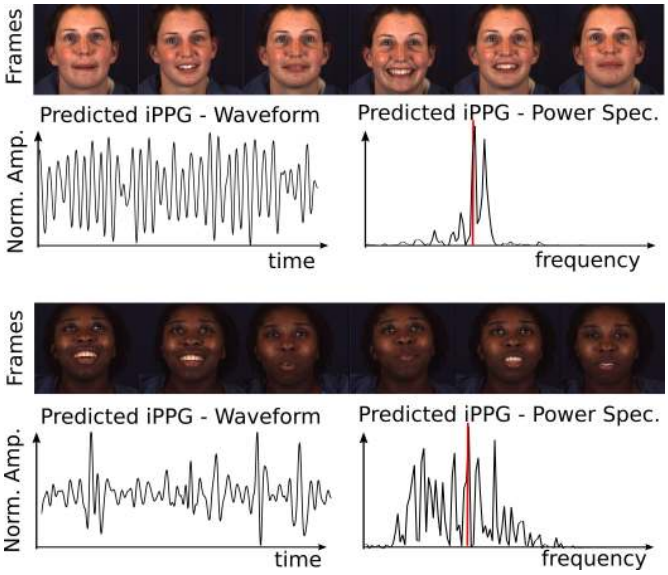


Figure 1. Characterizing the performance of computer vision algorithms is important to ensure that some demographic groups are not exposed to greater risk. We perform a meta-analysis of the impact of skin type on imaging photoplethysmography measurement. The examples shown illustrate that lighter skin types are associated with higher blood volume pulse signal-to-noise ratios. The red line indicates the heart rate measured via a contact sensor.

1. Introduction

There are inherent challenges with machine systems that rely on learning complex relationships from data, whether in a supervised or unsupervised manner. It is often difficult to be sure about how the resulting model will behave in practice. Specifically, regarding demographics, there is concern over differences in how algorithms perform on people from some groups compared to others [5]. Biases can be introduced in several parts of the model development process. If datasets are biased it can mean that performance is different and/or not carefully characterized across different groups [6, 7]. This is especially problematic if it leads to unjustified over-confidence in the technology. The data sources that provide training and testing examples in computer vision often have inherent biases [6, 8] as data from affluent, well-educated and over-represented populations are easier to collect. In recently published work, performance of facial gender classification algorithms has been shown to be worse for women and people with darker skin types [6, 7, 9]. Another study found that face detection performance suffered similarly and that the gender of people from East Asian countries was also more likely to be classified incorrectly [8]. Improving the performance of machine-learned classifiers is virtuous [10, 8]. However, an

Manuscript	N	M/F	Skin Types	Metrics	Results
Fallow et al. (2013) [1]	23	11 / 12	I/II-8, III-5, IV-4, V-6	BVP SNR	**BVP SNR lowest for the darkest skin types.**
Wang et al.(2015) [2]	15	N.R.	I/II-5, III-5, IV/V-5	BVP SNR, ρ, AUC	**BVP SNR lowest for the darkest skin types.** ρ comparable. No stats test performed.
Shao et al. (2016) [3]	23	15 / 8	II to V - Dist. unclear	BVP SNR	**BVP SNR lowest for the darkest skin types.**
Addison et al. (2018) [4]	10	5 / 5	I-2, II-2, III-2, IV-2, V-1, VI-1	HR bias, RMSD, ρ	**Errors highest for the darkest skin types.** No statistical tests performed.
Ours (2020)	83	45 / 38	I-3, II-25, III-24, IV-22, V-7, VI-3	BVP SNR, MAE, RMSE, ρ	Performance of supervised and unsupervised iPPG algorithms drops significantly for skin types V and VI.

Table 1. A summary of imaging PPG results by skin type from four previous studies and our meta-analysis. All studies have found their chosen metrics suffer on subjects with darker skin types. By performing a meta analysis we are able to analyze more than three times the number of subjects than any of the previous studies and perform statistical tests on the distributions of these results.

important first step is often becoming aware of the biases and why they occur (*"fairness through awareness"*) [11].

Imaging photoplethysmography (iPPG) is the name given to a set of techniques [12] for measuring the blood volume pulse via light reflected from the skin using imaging devices (e.g., a digital camera) [13, 14]. These methods have the potential to enable low-cost measurement of important vital signs, including: heart rate [15], blood oxygenation [16], respiration rate [17] and stress (as measured by changes in sympathetic nervous system activity) [18, 19]. Furthermore, even subtle dynamics of the morphology of the iPPG signal can be measured [18] - derivative metrics which have been linked to blood pressure [20, 21] and could be helpful in assessing the impact of chronic illnesses, such as hypertension, and the risk of mortality that come with those conditions.

Given that iPPG is an optical technique that relies on measuring light reflected from the skin it is natural to assume that skin type would influence the signal-to-noise ratio (SNR) and indeed several studies have found this to be the case [2, 4, 3, 1, 22]. A larger melanin concentration in people with darker skin absorbs more light, thus the intensity of light returning to the camera is lower and the iPPG signal weaker. For this work we use the Fitzpatrick Scale [23] for quantifying skin types, as this is the most commonly used. It has six skin type categories from I (lightest) to VI (darkest). With respect to gender there are less obvious reasons why one might hypothesize a difference in performance; however, women tend to have higher average resting heart rates than men [24]. One hypothesis might be that iPPG algorithms tuned on populations of mostly men may perform worse on videos of women (if they assume the heart rate should be present in a specific, lower, frequency range). Appearance differences between men and women including facial hair and the use of cosmetics [25], both of which can block light entering and leaving the skin, might impact the results too.

Only a few papers have reported iPPG results over different skin types [2, 4, 3, 1, 22] and all of these did so on their own private datasets which were limited in the number of subjects (in all cases 23 people or fewer). Due to the small number of subjects and lack of diversity most were not able to evaluate performance on all of the six Fitzpatrick skin types and to our knowledge none included a comparison across genders. To help address these gaps in the literature for this very important issue we perform a meta-analysis of results across three datasets and find that the blood volume pulse signal-to-noise ratio (BVP SNR) and the accuracy of heart rate measurement drop significantly for the darkest skin types (V an VI on the Fitzpatrick scale). We compare supervised and unsupervised learning algorithms and find that some algorithms are more robust to skin type parameters than others.

2. Representation in Computer Vision Datasets

Biases can occur for many reasons. "Selection bias" is related to the tendency for certain classes of data to be included in datasets in the first place. "Capture bias" is related to how the data are acquired and may be influenced by the collectors' preferences (e.g. in images of people with certain skin types or genders, etc.) and data collection settings. "Negative set bias" is related to the examples in the dataset which are supposed to represent the "rest of the world" [29] as negative instances of certain classes. In this paper, the

Table 2. Fitzpatrick skin type and gender distribution in three datasets (MMSE-HR [26], AFRL [27], MR-NIRP [28]). The number of subjects (and videos) is shown alongside example frames that illustrate the differences between skin types.

Fitzpatrick Skin Types **Gender**

Data	I	II	III	IV	V	VI	Female	Male	Total
[26]	0 / 0	8 / 22	11 / 26	18 / 44	2 / 6	2 / 4	23 / 64	17 / 38	40 / 102
[27]	1 / 12	13 / 156	10 / 120	1 / 12	0 / 0	0 / 0	8 / 96	17 / 204	25 / 300
[28]	0 / 0	2 / 3	1 / 2	1 / 2	4 / 8	0 / 0	2 / 4	6 / 11	8 / 15
Total	1 / 12	23 / 181	22 / 148	20 / 58	6 / 14	2 / 4	33 / 164	40 / 253	73 / 417

"capture bias" is the most relevant to our discussion as we consider the performance of iPPG methods on typically underrepresented groups.

For the broader context in computer vision, let us take several benchmark datasets used for facial analysis generally. Almost 50% of the people featured in the widely used MS-CELEB-1M dataset [30] are from North America and Western Europe, and over 75% are men. The demographic make up of these countries is predominantly lighter skin types.[1] Another dataset of faces, IMDB-WIKI [31], features 59.3% men and Americans are hugely over-represented (34.5%). Boulimwini et al. [6] found that the IARPA Janus Benchmark A (IJB-A) [32] contained only 7.80% of faces with skin types V or VI and again three-quarters of the participants were male. The datasets typically used for iPPG analyses (AFRL [27], MAHNOB [33], MMSE-HR [26] and VIPL [34]) have similar problems and have many fewer subjects. Thus, any one dataset may only have a few skin types represented and the distribution is usually biases towards types II and III. Moreover, because the iPPG is such a subtle intensity signal, it is easily corrupted by various sources of noise, including head motion, uncontrolled ambient illumination [28, 35, 22] and video quality [36, 37]. Therefore, it may be difficult to decouple the detrimental effects of motion and illumination noise on the quality of the iPPG signals from the effects of the demographics of the subjects in different datasets. Some datasets which contain more subjects of darker skin types may also contain different types of motion noise or be more compressed, leading to worse performance. These possible confounding sources of noise make it challenging to accurately characterize the effects of skin types on iPPG signal quality with a lack of sufficiently large and diverse publicly available datasets. Performing a meta-analysis helps us to increase the sample size in each skin type category and perform more robust statistical tests.

[1] http://data.un.org/

3. Imaging Photoplethysmography

Over the past 15 years iPPG has become an established research domain with applications in infant monitoring [38], telemedicine and affective computing [39]. As this technology begins to be deployed in real-world applications, it is imperative that the performance is carefully characterized. Table 1 provides a summary of research in which skin type analysis was performed to some extent. Fallow et al. [1] performed an analysis of pulse signal strength using a contact PPG device. Grouping the results into four categories (I/II, III, IV, V) they found that people with skin type V showed significantly lower BVP modulation. Wang et al. [2] performed a comparison of unsupervised iPPG methods and grouped the skin types of the 15 participants into three categories (I/II, III and IV/V). Again, they found that BVP SNR was lowest for the IV/V skin type group. Shao et al. [3] did not group skin type categories but only had subjects from type II to V (i.e., not the full range of skin types) again finding BVP SNR was lower in darker skin types. To our knowledge the only prior work to report results on subjects of all skin types is that of Addison et al. [4]. Of the 10 subjects in their study two were of each of the seven skin type categories, with the exception of categories V and VI that only had one subject. As with the other studies, they found that the performance dropped on subjects in skin type categories V and VI.

4. Experiments

4.1. Data

We perform a meta-analysis of three datasets to quantify how the iPPG signals are affected by different skin types and gender. Table 2 shows a summary of these three datasets and the distribution of the subjects with different skin types and genders. Examples of the video frames from the three datasets used are shown in Figure 2. In total we considered 417 videos of 73 subjects, featuring subjects from all skin type categories.

MMSE-HR [26] 102 videos of 40 participants were

Figure 2. Examples of video frames from datasets used to evaluate the effects of skin types: (a)AFRL[27], (b) MMSE-HR [26], (c) MR-NIRP (RGB) [28].

recorded at 25 frames per second (fps) and 1040x1392 pixel resolution during spontaneous emotion elicitation experiments. A gold-standard contact blood pressure wave was measured at 1000 fps. We computed heart rate estimates as the inverse of the time interval between the detected peaks of the blood pressure wave.

AFRL [27] 12 videos of 25 subjects were recorded at 120 fps at 658x492 pixel resolution during controlled head rotation tasks of varying speed of the motion of the head. The camera used was a color Scout scA640-120gc GigE-standard capturing 8-bit images. The six experiments involved the following motion tasks: 1) sitting still and resting the chin on a headrest, 2) sitting still without the headrests to allow for small natural head motion, 3) moving the head horizontally at a speed of 10 degrees/second, 4) 20 degrees/second, 5) 30 degrees/second, 6) reorienting the head randomly once every second to one of the nine predefined locations in a semicircle in front of the participant. Fingertip reflectance photoplethysmograms and electrocardiograms were recorded as gold-standard signals. We used the provided electrocardiogram signals to compute the final HR estimation errors.

MR-NIRP (RGB and NIR) [28] 15 videos of eight subjects were recorded, once during a stationary task, and once during a motion task involving talking and rigidly moving the head. All videos were simultaneously recorded with RGB (FLIR Blackfly BFLY-U3-23S6C-C) and near-infrared (NIR, Point Grey Grasshopper GS3-U3-41C6NIR-C) cameras. Raw images were captured with 640×640 resolution (10-bit depth) at 30 fps with fixed exposure, gamma correction turned off and gain set to zero. The ground-truth PPG waveform was recorded using a CMS 50D+ finger pulse oximeter at 60 fps. In this work we are only considering RGB camera recordings and therefore we do not use the NIR videos.

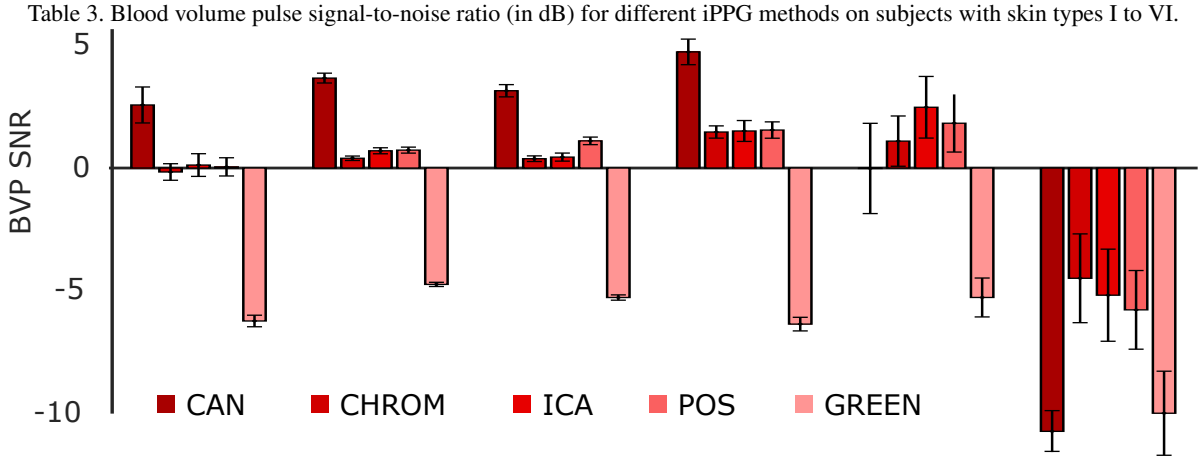
4.2. Methods

We do not necessarily expect all iPPG approaches to be impacted equally be skin type. We compare the performance of simply spatially averaging and filtering the green channel (GREEN [14]) camera signal with three state-of-the-art unsupervised methods (ICA [15], CHROM [35], POS [22]) which use heuristic physiological models for recovering the BVP, and the current state-of-the-art deep learning approach based on convolutional attention networks (CAN) [40]. Open source implementations of these approaches can be found in [41].

We trained the CAN model on the largest of the three datasets (AFRL [27]) using the provided contact pulse oximeter recordings as training labels. We trained the models separately on each of the six motion tasks from the AFRL dataset with a participant-independent cross-validation, leaving out 20% of the participants in each validation split. The MMSE-HR and MR-NIRP datasets are much smaller than AFRL and are not suited for training the complex CAN model. Therefore, for these datasets we used the model trained on AFRL motion Task 2. The participants in MMSE-HR and MR-NIRP did not move their heads out-of-plane, but had motions caused by facial expressions and talking. Therefore, Task 2 from the AFRL dataset was deemed the most similar to the motion present in the MMSE-HR and MR-NIRP datasets. To improve the generalizability to the new datasets we used a subject-dependent cross-validation to maximize the diversity of the participants that the CAN model was trained on, using the last 4 minutes of each video for training and the first 1 minute for testing.

In all cases the recovered BVP signal was filtered using a sixth-order Butterworth bandpass filter with cut-off frequencies of 0.7 and 2.5 Hz. The BVP signals were also preprocessed for each method with an AC/DC normalization by subtracting the pixel norm and dividing by the pixel standard deviation. The heart rate was estimated by computing the Fast Fourier Transforms (FFTs) of the estimated BVP signals and finding the frequency with the largest power spectrum energy in the range 0.7 and 2.5 Hz.

Table 3. Blood volume pulse signal-to-noise ratio (in dB) for different iPPG methods on subjects with skin types I to VI.



	Fitzpatrick Skin Types					
Method	I	II	III	IV	V	VI
CAN [40]	2.48 ± 2.24	3.54 ± 0.58	3.04 ± 0.70	4.57 ± 0.90	-0.02 ± 3.00	-10.35 ± 0.91
CHROM [35]	-0.16 ± 0.33	0.38 ± 0.09	0.37 ± 0.11	1.42 ± 0.24	1.06 ± 0.99	-4.34 ± 1.75
ICA [15]	0.12 ± 0.45	0.68 ± 0.12	0.43 ± 0.16	1.46 ± 0.41	2.39 ± 1.21	-5.00 ± 1.81
POS [22]	0.05 ± 0.36	0.70 ± 0.12	1.07 ± 0.15	1.49 ± 0.32	1.76 ± 1.13	-5.58 ± 1.55
GREEN [14]	-6.01 ± 0.23	-4.58 ± 0.08	-5.09 ± 0.1	-6.14 ± 0.27	-5.09 ± 0.76	-9.64 ± 1.65

5. Results

Comparison Across Skin Types. First, we perform a meta-analysis of the BVP SNR across the three datasets by skin type. Table 3 shows the SNR for each skin type. A bar chart reflects the numbers in the table. While the performance is relatively similar for skin types I to V, the BVP SNR drops dramatically for skin type VI. Therefore, while darker skin types lead to BVP signals with weaker amplitudes and are more prone to errors induced by different sources of noise, this most dramatically affects skin type VI and has only a modest effect on other skin types.

In addition to SNR, we compared the performance of three other error measures (heart rate MAE, RMSE, correlation). All of these measures capture the performance of heart rate estimation from the BVP. We show this comparison for the CHROM method for different skin types (the trends in results were similar for all methods). The effect of the reduction in BVP SNR is a severe drop in the performance of HR estimates (see Table 4), the MAE and RMSE increase three-fold for skin type category VI compared to category V.

Comparison Across Genders. The differences in the results for men and women are much less dramatic than the differences across skin types. Overall, the videos of females had marginally lower BVP SNR (men: 0.80, women: -0.35) and higher HR MAE (men: 3.78, women: 4.49) (see Table 5). All four metrics MAE, RMSE, SNR and ρ were slightly worse on videos of females than males (see Table 6).

Table 4. HR MAE, RMSE and ρ for the CHROM method on subjects with skin types I to VI.

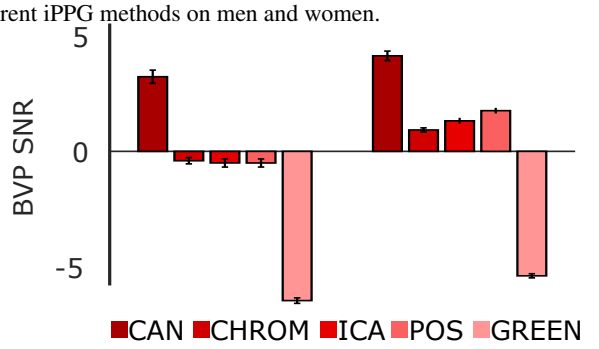
	Fitzpatrick Skin Types					
	I	II	III	IV	V	VI
MAE	3.42	4.09	4.23	2.14	3.23	13.58
RMSE	6.69	9.86	10.03	5.21	6.41	18.98
Corr., ρ	0.31	0.71	0.73	0.81	0.80	0.76

5.1. Comparison of Methods

The GREEN channel method performs poorly across all skin types mainly because it is not very robust to head motions or facial expressions. Each of the skin type categories had videos with significant motion which likely caused large errors in the GREEN method signal estimates. It is difficult to conclude much about how skin type affects this method as all the results are poor.

The supervised neural network (CAN) model performed significantly better than all unsupervised methods on skin types II to IV. These are the skin types which are predominantly present in the AFRL dataset that the CAN was trained on. On darker skin types V and VI all methods, including CAN perform very poorly, with CAN performing worse than all of the unsupervised methods. Therefore, even though deep learning models may be able to learn to robustly recover the BVP signals in diverse noise scenarios and could become invariant to darker skin types, a model that was not trained on such videos may perform significantly worse. This is a sign that the supervised model over-

Table 5. Blood volume pulse signal-to-noise ratio (in db) for different iPPG methods on men and women.



	Gender	
Method	Female	Male
CAN [40]	2.78 ± 0.62	3.55 ± 0.51
CHROM [35]	-0.35 ± 0.12	0.80 ± 0.08
ICA [15]	-0.43 ± 0.15	1.14 ± 0.11
POS [22]	-0.43 ± 0.15	1.51 ± 0.10
GREEN [14]	-5.55 ± 0.10	-4.63 ± 0.08

Table 6. HR MAE, RMSE and ρ for the CHROM method on men and women.

	Gender	
	Female	Male
MAE	4.49	3.78
RMSE	10.44	9.17
Corr., ρ	0.63	0.76

fits to the training corpus. There is some hope that with more data of participants with skin type VI this problem could be partially rectified.

The unsupervised methods also perform best on intermediate skin types (III and IV) but performance drops for all methods (CHROM, ICA, POS) on type VI. The CHROM method seems to show the most robust (similar performance) across all skin types, followed by the POS method.

6. Discussion

For this meta-analysis we combined three datasets, each with different participants, behaviors (tasks) and captured in different conditions (e.g., hardware and lighting conditions). We find that skin types impact iPPG measurement, but most significantly for the darkest skin type (VI). Performance differences across all other skin types were only small.

Other factors (including motion) have a significant impact on the results, perhaps even more so than skin type. Even though the motion tasks within each dataset for different videos and participants were similar, it is possible that some videos have larger and more challenging motion than others in the same skin type category. This is partly captured by the relatively large error-bars, especially for skin types in which the number of subjects is comparatively small (e.g., I and VI).

There are several limitations that should be acknowledged with these analyses. First, even after combining three datasets, we only have one participant of skin type I and two of skin type VI, making it difficult to draw highly generalizable conclusions. It is crucial to continue to study this issue as larger and more diverse datasets become publicly available. Second, a similar problem of worse performance on darker skin types might also occur with contact devices, as shown by Fallow et al. [1]. BVP signals with lower amplitude caused by darker skin types could be more prone to error artifacts. This could possibly lead to erroneous gold-standard measurements, making it harder to validate video-based iPPG methods.

Using NIR recordings could be a potential solution, as NIR is robust to darker skin types because the NIR light penetrates deeper into the skin. Currently, there is no public dataset with NIR recordings and ground truth PPG signals with a sufficiently large number and diverse participants to validate this claim. Furthermore, the BVP SNR is weaker in general in the NIR range compared to RGB [28], making it less motion-robust.

7. Conclusions

It is important that machine vision systems, especially those that may be used in medical or health related applications, are well validated and performance is characterized to ensure that certain demographic groups are not exposed to greater risks than others. To our knowledge, this is the first meta-analysis of the impact of skin type on the performance of iPPG systems. We find that the performance of heart rate estimation from video is significantly worse for people with skin types in category VI of the Fitzpatrick scale. Across skin types I to V performance is relatively stable. We find that there are slight differences in performance across genders with videos of females tending to have lower accuracy, but the differences are not significant. The community should focus greater efforts on developing methods that are more robust to differences in skin types. To this end, datasets with better representation of all skin types would be very helpful.

References

[1] Bennett A Fallow, Takashi Tarumi, and Hirofumi Tanaka. Influence of skin type and wavelength on light wave reflectance. *Journal of clinical monitoring and computing*, 27(3):313–317, 2013. 2, 3, 6

[2] Wenjin Wang, Sander Stuijk, and Gerard De Haan. A novel algorithm for remote photoplethysmography: Spatial subspace rotation. *IEEE transactions on biomedical engineering*, 63(9):1974–1984, 2015. 2, 3

[3] Dangdang Shao, Francis Tsow, Chenbin Liu, Yuting Yang, and Nongjian Tao. Simultaneous monitoring of ballistocardiogram and photoplethysmogram using a camera. *IEEE Transactions on Biomedical Engineering*, 64(5):1003–1010, 2016. 2, 3

[4] Paul S Addison, Dominique Jacquel, David MH Foo, and Ulf R Borg. Video-based heart rate monitoring across a range of skin pigmentations during an acute hypoxic challenge. *Journal of clinical monitoring and computing*, 32(5):871–880, 2018. 2, 3

[5] Executive Office of the President, Cecilia Munoz, Domestic Policy Council Director, Megan (US Chief Technology Officer Smith (Office of Science, Technology Policy)), DJ (Deputy Chief Technology Officer for Data Policy, Chief Data Scientist Patil (Office of Science, and Technology Policy)). *Big data: A report on algorithmic systems, opportunity, and civil rights*. Executive Office of the President, 2016. 1

[6] Joy Adowaa Buolamwini. *Gender shades: intersectional phenotypic and demographic evaluation of face datasets and gender classifiers*. PhD thesis, Massachusetts Institute of Technology, 2017. 1, 3

[7] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, pages 77–91, 2018. 1

[8] Daniel McDuff, Shuang Ma, Yale Song, and Ashish Kapoor. Characterizing bias in classifiers using generative models. In *Advances in Neural Information Processing Systems*, pages 5404–5415, 2019. 1

[9] Inioluwa Deborah Raji and Joy Buolamwini. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In *AAAI/ACM Conf. on AI Ethics and Society*, 2019. 1

[10] Hee Jung Ryu, Margaret Mitchell, and Hartwig Adam. Improving smiling detection with race and gender diversity. *arXiv preprint arXiv:1712.00193*, 2017. 1

[11] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226. ACM, 2012. 2

[12] Daniel McDuff, Justin R Estepp, Alyssa M Piasecki, and Ethan B Blackford. A survey of remote optical photoplethysmographic imaging methods. In *2015 37th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, pages 6398–6404. IEEE, 2015. 2

[13] Chihiro Takano and Yuji Ohta. Heart rate measurement based on a time-lapse image. *Medical engineering & physics*, 29(8):853–857, 2007. 2

[14] Wim Verkruysse, Lars O Svaasand, and J Stuart Nelson. Remote plethysmographic imaging using ambient light. *Optics express*, 16(26):21434–21445, 2008. 2, 4, 5, 6

[15] Ming-Zher Poh, Daniel McDuff, and Rosalind W Picard. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Optics express*, 18(10):10762–10774, 2010. 2, 4, 5, 6

[16] L Tarassenko, M Villarroel, A Guazzi, J Jorge, DA Clifton, and C Pugh. Non-contact video-based vital sign monitoring using ambient light and auto-regressive models. *Physiological measurement*, 35(5):807, 2014. 2

[17] Ming-Zher Poh, Daniel McDuff, and Rosalind W Picard. Advancements in noncontact, multiparameter physiological measurements using a webcam. *IEEE transactions on biomedical engineering*, 58(1):7–11, 2010. 2

[18] Daniel McDuff, Sarah Gontarek, and Rosalind Picard. Remote measurement of cognitive stress via heart rate variability. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 2957–2960. IEEE, 2014. 2

[19] Izumi Nishidate, Chihiro Tanabe, Daniel J McDuff, Kazuya Nakano, Kyuichi Niizeki, Yoshihisa Aizu, and Hideaki Haneishi. Rgb camera-based noncontact imaging of plethysmogram and spontaneous low-frequency oscillation in skin perfusion before and during psychological stress. In *Optical Diagnostics and Sensing XIX: Toward Point-of-Care Diagnostics*, volume 10885, page 1088507. International Society for Optics and Photonics, 2019. 2

[20] Paul S Addison. Slope transit time (stt): A pulse transit time proxy requiring only a single signal fiducial point. *IEEE Transactions on Biomedical Engineering*, 63(11):2441–2444, 2016. 2

[21] Mohamed Elgendi, Richard Fletcher, Yongbo Liang, Newton Howard, Nigel H Lovell, Derek Abbott, Kenneth Lim, and Rabab Ward. The use of photoplethysmography for assessing hypertension. *NPJ digital medicine*, 2(1):1–11, 2019. 2

[22] 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. 2, 3, 4, 5, 6

[23] Thomas B Fitzpatrick. The validity and practicality of sun-reactive skin types i through vi. *Archives of dermatology*, 124(6):869–871, 1988. 2

[24] Ken Umetani, Donald H Singer, Rollin McCraty, and Mike Atkinson. Twenty-four hour time domain heart rate variability and heart rate: relations to age and gender over nine decades. *Journal of the American College of Cardiology*, 31(3):593–601, 1998. 2

[25] Wenjin Wang and Caifeng Shan. Impact of makeup on remote-ppg monitoring. *Biomedical Physics & Engineering Express*, 6(3):035004, 2020. 2

[26] Zheng Zhang, Jeff M Girard, Yue Wu, Xing Zhang, Peng Liu, Umur Ciftci, Shaun Canavan, Michael Reale, Andy Horowitz, Huiyuan Yang, et al. Multimodal spontaneous emotion corpus for human behavior analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3438–3446, 2016. 3, 4

[27] Justin R Estepp, Ethan B Blackford, and Christopher M Meier. Recovering pulse rate during motion artifact with a multi-imager array for non-contact imaging photoplethysmography. In *2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 1462–1469. IEEE, 2014. 3, 4

[28] Ewa Magdalena Nowara, Tim K Marks, Hassan Mansour, and Ashok Veeraraghavan. Sparseppg: towards driver monitoring using camera-based vital signs estimation in near-infrared. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1353–135309. IEEE, 2018. 3, 4, 6

[29] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1521–1528. IEEE, 2011. 2

[30] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European Conference on Computer Vision*, pages 87–102. Springer, 2016. 3

[31] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Dex: Deep expectation of apparent age from a single image. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 10–15, 2015. 3

[32] Brendan F Klare, Ben Klein, Emma Taborsky, Austin Blanton, Jordan Cheney, Kristen Allen, Patrick Grother, Alan Mah, and Anil K Jain. Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1931–1939, 2015. 3

[33] Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic. A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing*, 3(1):42–55, 2011. 3

[34] Xuesong Niu, Hu Han, Shiguang Shan, and Xilin Chen. Vipl-hr: A multi-modal database for pulse estimation from less-constrained face video. In *Asian Conference on Computer Vision*, pages 562–576. Springer, 2018. 3

[35] Gerard De Haan and Vincent Jeanne. Robust pulse rate from chrominance-based rppg. *IEEE Transactions on Biomedical Engineering*, 60(10):2878–2886, 2013. 3, 4, 5, 6

[36] Ewa Nowara and Daniel McDuff. Combating the impact of video compression on non-contact vital sign measurement using supervised learning. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019. 3

[37] Zitong Yu, Wei Peng, Xiaobai Li, Xiaopeng Hong, and Guoying Zhao. Remote heart rate measurement from highly compressed facial videos: an end-to-end deep learning solution with video enhancement. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 151–160, 2019. 3

[38] Mauricio Villarroel, Sitthichok Chaichulee, João Jorge, Sara Davis, Gabrielle Green, Carlos Arteta, Andrew Zisserman, Kenny McCormick, Peter Watkinson, and Lionel Tarassenko. Non-contact physiological monitoring of preterm infants in the neonatal intensive care unit. *npj Digital Medicine*, 2(1):1–18, 2019. 3

[39] Daniel J McDuff, Javier Hernandez, Sarah Gontarek, and Rosalind W Picard. Cogcam: Contact-free measurement of cognitive stress during computer tasks with a digital camera. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 4000–4004, 2016. 3

[40] Weixuan Chen and Daniel McDuff. Deepphys: Video-based physiological measurement using convolutional attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 349–365, 2018. 4, 5, 6

[41] Daniel McDuff and Ethan Blackford. iphys: An open non-contact imaging-based physiological measurement toolbox. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 6521–6524. IEEE, 2019. 4