Remote Blood Oxygen Estimation From Videos Using Neural Networks

Joshua Mathew
NC State University
jrmathew@ncsu.edu

Xin Tian
Univeristy of Maryland
xtian17@umd.edu

Min Wu
Univeristy of Maryland
minwu@umd.edu

Chau-Wai Wong
NC State University
chauwai.wong@ncsu.edu

Abstract

Blood oxygen saturation (SpO_2) is an essential indicator of respiratory functionality and is receiving increasing attention during the COVID-19 pandemic. Clinical findings show that it is possible for COVID-19 patients to have significantly low SpO_2 before any obvious symptoms. The prevalence of cameras has motivated researchers to investigate methods for monitoring SpO_2 using videos. Most prior schemes involving smartphones are contact-based: They require a fingertip to cover the phone's camera and the nearby light source to capture re-emitted light from the illuminated tissue. In this paper, we propose the first convolutional neural network based noncontact SpO_2 estimation scheme using smartphone cameras. The scheme analyzes the videos of a participant's hand for physiological sensing, which is convenient and comfortable, and can protect their privacy and allow for keeping face masks on. We design our neural network architectures inspired by the optophysiological models for SpO_2 measurement and demonstrate the explainability by visualizing the weights for channel combination. Our proposed models outperform the state-of-the-art model that is designed for contact-based SpO_2 measurement, showing the potential of our proposed method to contribute to public health. We also analyze the impact of skin type and the side of a hand on SpO_2 estimation performance.

1. Introduction

Blood circulation underneath a person's skin induces subtle color variations of the skin area. These subtle changes can be captured without contact using ubiquitous cameras, such as low-cost webcams and smartphone cameras. Such noncontact video-based measurement contains important health-related information of a person and provides an agile way to monitor health vital signs, including heart rate [9, 17, 32, 37], breathing rate [6, 23], and heart rate variability [14, 23, 11], as an emerging class of biometrics enabled by computer vision techniques.

Blood oxygen saturation (SpO_2) is an important physiological parameter that represents the level of oxygen supply in the blood and reflects the adequacy of respiratory function [20]. The estimation and monitoring of SpO_2 are essential for the assessment of lung function and the treatment of chronic pulmonary diseases, especially during the COVID-19 pandemic when it has been reported that patients being infected by the virus can have significantly low SpO_2 before any obvious respiratory symptoms occur [8, 29]. The conventional SpO_2 measurement methods rely on contact-based sensing, including fingertip pulse oximetry and its variants in smartwatches and smartphones [25, 24, 18, 10]. These conventional contact-based methods may cause discomfort and skin irritation, especially for people with sensitive skin, and are not always accessible to the public [29]. In recent years, a growing number of studies have investigated SpO_2 measurement using videos [15, 33, 26, 31, 34, 3, 5, 30], which allows for SpO_2 estimation without contact. These video-based noncontact methods provide a more comfortable and unobtrusive way to monitor SpO_2, and have the potential to be adopted in health screening and telehealth.

Based on the setup of cameras and light sources, existing noncontact, video-based SpO_2 estimation methods can be grouped into two main categories. Methods from the first category utilize monochromatic sensing similar to the conventional pulse oximetry. They use either high-end monochromatic cameras with selected optical filters or controlled monochromatic light sources [15, 33, 26, 31]. The other category uses consumer-grade RGB cameras, such as digital webcams [30, 3, 5].

All these video-based noncontact SpO_2 estimation methods utilize the differences in the optophysiological characteristics of oxygenated hemoglobin and deoxygenated hemoglobin. The monochromatic light sources and sensors are selected to have accurate control of the absorption effect

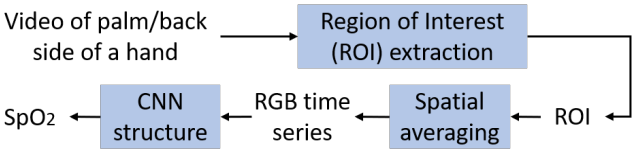


Figure 1: Proposed SpO_2 estimation method. Three color time series are extracted from the skin area of a hand video, and are then fed into an optophysiology-inspired neural network for SpO_2 prediction.

of hemoglobin, while the consumer-grade digital cameras, including webcams and smartphone cameras, have a wider sensing band and are more challenging for SpO_2 sensing.

Inspired by the optophysiological model [36, 33, 24], in this paper, we investigate the design of explainable neural networks for extracting features from a video stream using consumer-grade cameras capturing a participant's hand to monitor SpO_2. To the best of our knowledge, there is no prior work that remotely monitors SpO_2 with regular RGB cameras using neural networks.

In this paper, we propose using convolutional neural networks (CNN) for contactless SpO_2 monitoring from videos captured by smartphone cameras. Fig. 1 is an overview of the system design. First, the region of interest (ROI), including the palm and back side of the hand, is extracted from the smartphone captured videos. Second, the ROI is spatially averaged to produce R, G, and B time series. Third, the three time series are fed into an optophysiology-inspired CNN for SpO_2 estimation. We consider the hand region in this work as a proof-of-concept. Compared to using the face for SpO_2 measurement as most of the prior art did [3, 30], recording hand videos raises less privacy concern to the participants and is a safer way for data collection during the COVID-19 pandemic according to the mask wearing guideline of human subject research. The contributions of our work are summarized as follows:

- This is the first work to use neural networks to address the challenging problem of contactless SpO_2 sensing using consumer-grade RGB cameras.
- Through a data-driven approach and visualization of the weights for the RGB channel combinations, we demonstrate the explainability of our model and that the choice of the color band learned by the neural network is consistent with the suggested color bands used in the optophysiological methods.
- We analyze the impact of the two sides of the hand and different skin tones on the quality of SpO_2 estimation.
- We achieve more accurate SpO_2 estimation with our optophysiologically inspired neural network structures when compared to the state-of-the-art neural network structure for contact-based SpO_2 prediction.

2. Background and Related Work

Optophysiological Model for Blood Oxygen Saturation Measurement. The protein molecule hemoglobin (Hb) in the blood carries oxygen from the lungs to tissues of the body. The level of blood oxygen saturation (SpO_2) reflects the ratio of oxygenated hemoglobin (HbO_2) to total hemoglobin and indicates the adequacy of respiratory function [20]. The normal range of SpO_2 is 95% to 100% [20]. Abnormality in the SpO_2 level can serve as an early warning sign of respiratory diseases [20]. A convenient and noninvasive way to continuously measure SpO_2 is pulse oximetry [25]. Pulse oximeters utilize the *principle of ratio of ratios* that was first proposed by Aoyagi in the early 1970s [25], and nowadays pulse oximeters are commonly used in hospitals, clinics, and homes. The ratio-of-ratios method leverages the optical absorbance difference of Hb and HbO_2 at two wavelengths, conventionally, at red and infrared wavelengths as indicated on Fig. 2. For the commonly seen pulse oximeters, lights at the red and infrared wavelengths are emitted through the fingertip. The transmitted light, interacted and attenuated by the blood and tissue, and received by an optical sensor, conveys information about pulsatile blood volume. The pulsatile blood volume at the two wavelengths is further processed to obtain an SpO_2 estimate.

With the prevalence of smartphones, researchers have investigated methods of monitoring SpO_2 using smartphones, most of which are contact-based and require the fingertips to be pushed against the illuminated light source and the built-in camera [24, 18, 10], so that the diffusely reflected light by the fingertip is captured by the camera. In this setup, an adapted ratio-of-ratios model is utilized with the red and blue channels of color videos in lieu of the traditional narrowband red and infrared wavelengths.

Deep Learning Aided Camera-based Physiological Monitoring. Deep learning has demonstrated promising performance in camera-based physiological measurement, such as heart rate and breathing rate [21, 7, 28]. An end-to-end convolutional attention network was proposed in [7] to estimate the blood volume pulse from face videos. Frequency analysis is then conducted on the estimated pulse signal for heart rate and breathing rate tracking. The study in [21] demonstrates that the heart rate can be directly inferred using a convolutional network with spatial-temporal representations of the face videos as its input.

Deep learning for SpO_2 monitoring from videos is still in its early stage. Ding et al. [10] proposed a convolutional neural network architecture for contact-based SpO_2 monitoring with smartphone cameras. Even though they have shown better performance than the conventional ratio-of-ratios method, their technique requires the users' fingertips

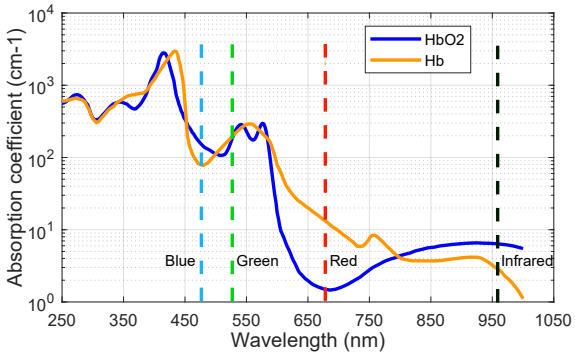


Figure 2: Extinction coefficient curves of hemoglobin. The curves were plotted based on [10, 1]. The difference between oxygenated hemoglobin (HbO_2) and deoxygenated hemoglobin (Hb) at the red and blue wavelengths means that these color channels contain useful information for SpO_2 prediction by means of optophysiological principles.

to be in contact with the illuminated flashlight and camera, which may not only lead to a sense of burning for a continuous period of time but also raise sanitation concerns, especially if the sensing device is reused by different participants. This motivates us to develop a deep learning architecture to monitor SpO_2 in a contactless way with regular RGB cameras, which has the potential to be adopted in health screening and telehealth.

3. Proposed Method for SpO_2 From Videos

We aim to estimate SpO_2 levels using a hand video by leveraging the fact that the color of the skin changes subtly when red cells in the blood carry/release oxygen. In our proposed method, we extract three color time series from the skin area of the hand video. We feed the extracted time series to optophysiology-inspired neural networks designed to achieve better and more explainable SpO_2 predictions.

3.1. Extraction of Skin Color Signals

The physiological information related to SpO_2 is embedded in the color of the reflected/reemitted light from a person's skin. Hence, a preprocessing step that precisely extracts the color information from the skin area is crucial to the design of an effective SpO_2 estimation method. For each participant's video, we aim to extract the R, G, and B time series and refer to these 1-D time series as *skin color signals*. We first need to locate the ROI of the skin pixels from the video. We found that it is most effective to discriminate the skin pixels from the background along the Cr color direction of the YCbCr color space [4]. We use Otsu's method [22] to determine a threshold that best separates the skin pixels from the background by minimizing the variance within the skin and non-skin classes. Once the ROI corresponding to the hand is located, the R, G, and B time series are generated by spatially averaging over the values of skin pixels for each frame of the video.

The skin color signals are split up into 10-second segments using a sliding window with a step size/stride of 0.2 seconds to serve as the inputs for neural networks. From a optophysiological perspective, the reflected/reemitted light from the skin for the duration of one cycle of heartbeat, i.e., 0.5–1 seconds for a heart rate of 60–120 bpm, should contain almost the complete information necessary to estimate the instantaneous SpO_2 [25]. In our system design, we use longer segments to add resilience against sensing noise. Since the segment length is one order of magnitude longer than the minimally required length to contain the SpO_2 information, we can use a fully-connected or convolutional structure to adequately capture the temporal dependencies without resorting to a recurrent neural network structure.

3.2. Neural Network Architectures

The previous neural network work for SpO_2 prediction mainly explored prediction, but not the model explainability [10]. Explainability/interpretability is highly desirable in many applications yet often not sufficiently addressed, partly due to the black box nature of neural networks. From a healthcare standpoint, explainability is a key factor which should be taken into account at the beginning of the design of a system. To extract features from the skin color signals and estimate SpO_2, we propose three physiologically motivated neural network structures. These structures are inspired by domain knowledge-driven physiological sensing methods and designed to be physically explainable. For heart rate sensing [37, 21] and respiratory rate sensing [19, 27], the RGB skin color signals are often combined first, as in the plane-orthogonal-to-skin (POS) algorithm [35], followed by temporal feature extraction. In contrast, for conventional SpO_2 sensing methods such as the ratio-of-ratios [36], the color components are combined at the end. Our proposed neural network structures explore different arrangements of channel combination and temporal feature extraction. We want to systematically compare the performance of our explainable model structures.

Channel Mixing Followed by Feature Extraction. In Model 1, shown as the leftmost structure depicted in Fig. 3, we combine the color channels first using several channel combination layers and then extract temporal features using temporal convolution and max pooling. A channel combination layer first linearly combines the C_{in} input channels/vectors into C_{out} activation vectors and then applies a rectified linear unit (ReLU) activation function to obtain the output channels/vectors. Mathematically, the channel combination layer is described as follows:

$$\mathbf{V} = \sigma(\mathbf{W}\mathbf{U} + \mathbf{b}\mathbb{1}^T), \tag{1}$$

where $\mathbf{U} \in \mathbb{R}^{C_{\text{in}} \times L}$ is the input comprised of C_{in} time series/vectors of length L. The initial channel combination

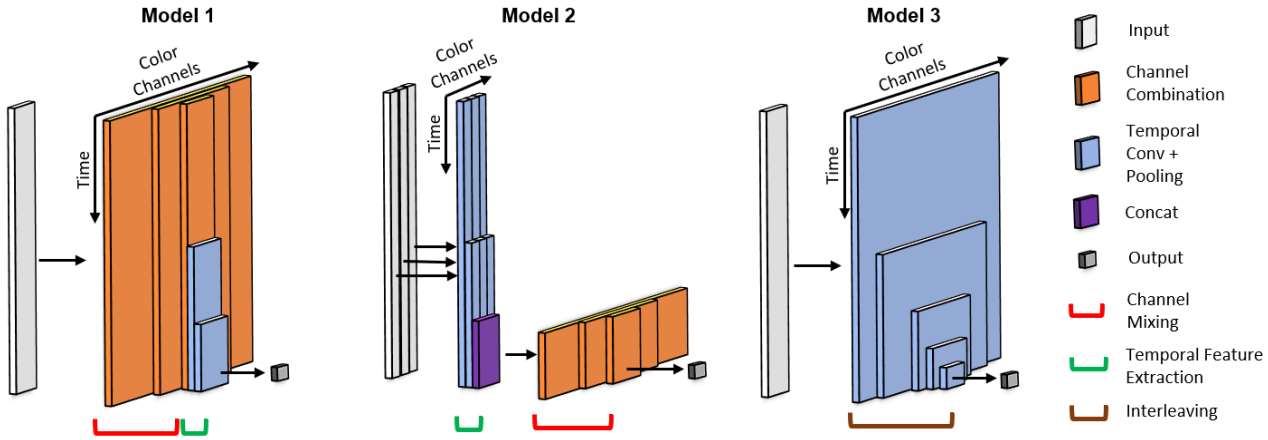


Figure 3: Proposed network structures for predicting an SpO_2 level from a fixed-length segment of skin color signals. We highlight the differences among three model configurations instead of showing the exact model structures. Model 1 combines the RGB channels before temporal feature extraction. Model 2 extracts the temporal features from each channel separately and fuses them toward the end. Model 3 interleaves color channel mixing and temporal feature extraction .

layer has an input of three channels with 300 points along the time axis. $\mathbf{W} \in \mathbb{R}^{C_{\text{out}} \times C_{\text{in}}}$ is a weight matrix, where each of the C_{out} rows of the matrix is a different linear combination for the input channels. A bias vector $\mathbf{b} \in \mathbb{R}^{C_{\text{out}}}$ contains the bias terms for each of the C_{out} output channels, which ensures that each data points in the artificially created segment of length L has the same intercept. $\mathbb{1}^T \in \mathbb{R}^{1 \times L}$ is a row vector of all ones. The nonlinear ReLU function $\sigma(x) = \max(0, x)$ is applied elementwise to the activation map/matrix. The output of the channel combination layer $\mathbf{V} \in \mathbb{R}^{C_{\text{out}} \times L}$ contains C_{out} channels of nonlinearly combined input channels.

The channel mixing section concatenates multiple channel combination layers with decreasing channel counts to provide significant nonlinearity. The output of the last channel combination layer has seven channels. After the channel mixing, for temporal feature extraction, we utilize multiple convolutional and max pooling layers with a downsampling factor of two to extract the temporal features of the channel-mixed signals. When there are multiple filters in the convolutional layer, then there will also be some additional channel combining with each filter outputting a channel-mixed signal. Finally, a single node is used to represent the predicted SpO_2 level.

Feature Extraction Followed by Channel Mixing. In Model 2, the middle structure depicted in Fig. 3, we reverse the order of channel mixing and temporal feature extraction from that in Model 1. The three color channels are separately fed for temporal feature extraction. The convolutional layers learn different features unique to each channel. At the output of the temporal feature extraction section, each color channel has been downsampled to retain only the important temporal information. The color channels are then mixed together in the same way as described for Model 1 before outputting the SpO_2 value.

Interleaving Feature Extraction and Channel Mixing. In our third model, we explore the possibility of interleaving the color channel mixing and temporal feature extraction steps. As illustrated by the rightmost structure depicted in Fig. 3, the input is first put through a convolutional layer with many filters and then passed to max pooling layers, resulting in feature extraction along the time as well channel combinations through each filter. The number of filters is reduced with each successive convolutional layer, gradually decreasing the number of combined channels and downsampling the signal in the time domain.

Loss Function and Parameter Tuning. We use the root-mean-squared-error (RMSE) as the loss function for all models. During training, we save the model instance at the epoch with the lowest validation loss. The neural network inputs are scaled to have zero mean and unit variance to improve the numerical stability of the learning. The parameters and hyperparameters of each model structure were tuned using the HyperBand algorithm [16] which allows for faster and more efficient search over a large parameter space than grid search or random search. It does this by running random paramater configurations on a specific schedule of iterations per configuration, and uses earlier results to select candidates for longer runs. The parameters that were tuned include the learning rate, the number of filters and kernel size for convolutional layers, the number of nodes, the dropout probability, and whether to do batch normalization after each convolutional layer.

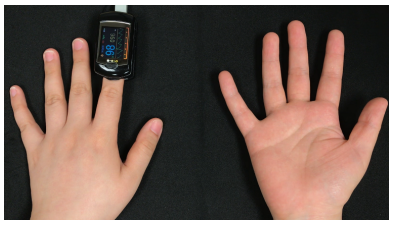
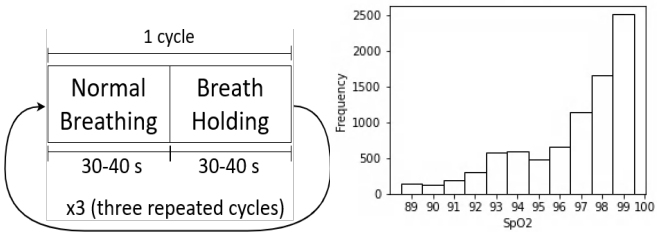
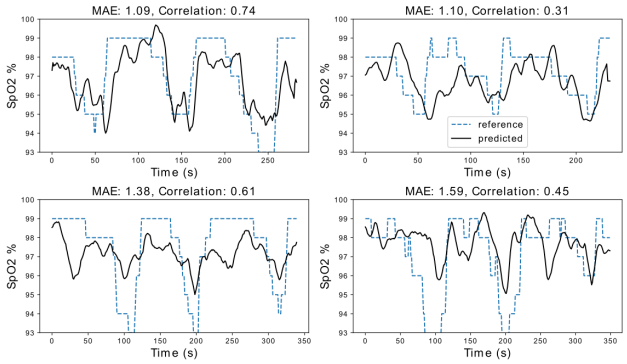


Figure 4: Illustration of two hand-video capturing positions. The hand on the left is in the palm down (PD) position and the hand on the right is in the palm up (PU) position.

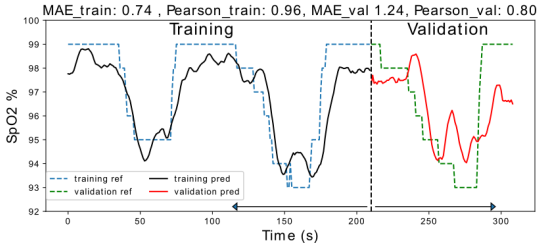


(a) (b)

Figure 5: (a) Breathing protocol that participants were asked to follow, including 3 cycles of normal breathing and breath holding. (b) Histogram of SpO_2 values in the collected dataset.



(a)



(b)

Figure 6: (a) Test predictions of varying performance with reference SpO_2. (b) Training vs. validation predictions. The higher the Pearson correlation, the better the predictions captures the reference SpO_2 trend. The lower the MAE, the better the predictions capture the dips in SpO_2.

4. Experimental Results

4.1. Dataset and Capturing Conditions

Our proposed models were evaluated on a self-collected dataset. The dataset consisted of hand video recordings and SpO_2 data from 14 participants, of which there were six males and eight females between the ages of 21 and 30. Participants were asked to categorize their skin tone based on the Fitzpatrick skin types [2] shown in Fig. 10. The distribution of the participants' skin types is as follows: Two participants of type II, eight participants of type III, one participant of type IV, and three participants of type V. This research was approved by the University of XXX Institutional Review Board.

Our dataset consists of four recordings per participant for a total of 56 recordings. Each participant was asked to place his/her hands on a table with the palm of the left hand and the back of the right hand facing the camera, as illustrated in Fig. 4. We refer to these two hand-video capturing positions as *palm up (PU)* and *palm down (PD)*, respectively. Each participant was asked to follow the breathing protocol outlined in Fig. 5a. The participant breathes normally for 30–40 seconds and then holds his/her breath for 30–40 seconds, and this process is repeated three times for each recording. All videos were recorded using an iPhone 7 Plus. The participant's SpO_2 was simultaneously measured using a Contec CMS50E pulse oximeter clamped to the left index finger of the hand. We use this pulse oximeter as the reference measurement as it has been validated to be within $\pm 2\%$ of the true SpO_2 level for the range of SpO_2 levels in our dataset. The video frame rate is 30 fps and the sampling rate for the reference SpO_2 measurements is 1 Hz. This data capturing procedure was repeated twice for each participant with at least 15 minutes between sessions.

The reference SpO_2 signal is interpolated to 5 sample points per second to match the segment sampling rate using a smooth spline approximation [12]. Each RGB segment and SpO_2 value pair is fed into our models as a single data point, the models output a single SpO_2 estimate per segment. To evaluate a model on a recording, the model is sequentially fed all RGB segments from the recording to generate a time series of preliminarily predicted SpO_2 values. All predictions greater than 100% SpO_2 are clipped to 100% since they are physiologically impossible. A 10-second long moving average filter is applied to generate a refined time series of predicted SpO_2 values.

4.2. Participant-Specific Results

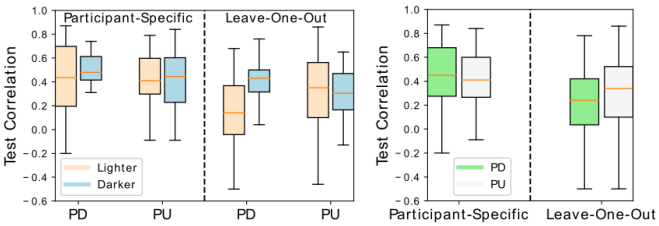
To investigate how well the proposed models could learn to estimate a specific individual's SpO_2 from his/her own data, we first conducted participant-specific experiments, that is, we learn individualized models for each participant.

Experimental Setting.

Two recordings per participant were captured with at

Table 1: Performance comparison of each model structure for participant-specific experiments. Results are given as the median of all participants.

	Hand Mode	Correlation			MAE (%)			RMSE (%)		
		Train	Val	Test	Train	Val	Test	Train	Val	Test
Model 1 (Proposed)	PD	0.86	0.75	0.41	1.90	1.52	2.12	2.26	1.94	2.51
	PU	0.78	0.82	0.39	1.32	1.26	2.16	1.54	1.60	2.70
Model 2 (Proposed)	PD	0.79	0.74	**0.46**	1.55	1.63	2.09	1.91	1.98	2.52
	PU	0.86	0.77	**0.41**	1.08	1.52	1.96	1.34	1.70	2.48
Model 3 (Proposed)	PD	0.81	0.77	0.44	1.64	1.27	**1.93**	1.99	1.59	**2.48**
	PU	0.93	0.80	**0.41**	1.50	1.25	**1.81**	1.72	1.47	**2.43**
Ding et al. [10]	PD	0.82	0.71	0.38	1.75	1.39	3.25	2.09	1.73	3.83
	PU	0.83	0.72	0.34	1.60	1.26	3.40	1.93	1.58	4.58



(a) (b)

Figure 7: Box plots comparing distributions of correlations for (a) lighter vs. darker skin types, and (b) PD vs. PU for all skin types. The PD results are better for darker skin tones in both the participant-specific and leave-one-out cases.

least 15 minutes in between. One recording is used for training and validation of the model and the remaining recording is for testing. An example of the training and validation predictions curves are shown in Fig. 6b. Each recording contains three breathing cycles, for each training/validation recording, the first two breathing cycles are taken for training and the third cycle is used for validation. Splitting the recordings into cycles instead of randomly sampling the 10-sec overlapping RGB segments ensures that there are no overlapping segments of data between the training and validation set. Example test prediction curves and their correlation and mean-absolute-error (MAE) are shown for reference in Fig. 6a. It should be noted that if the correlation is low, e.g., a constant temporal estimate, then the MAE and RMSE metrics are less meaningful. For the participant-specific experiments, due to the small dataset size, we augment the training and validation data by sampling with replacement. The oversampling also helps address the imbalance in SpO_2 data values that is shown in Fig. 5b.

In each experiment, the model structure and hyperparameters are first tuned using the training and validation data. Once the model has been tuned, we train multiple instances of the model using the best tuned hyperparameters. Between each instance, we vary the random seed used for model weights initialization and random oversampling. Each model instance is evaluated on the training/validation recording, the model instance that achieves the highest validation RMSE is selected for evaluation on the test recording. This model is then evaluated on the test recording to obtain the final test results.

Results. Table 1 shows the performance comparison of our proposed models with the prior-art model from Ding et al. [10]. To the best of our knowledge, this is the only convolutional neural network structure that has been tried for this same task of SpO_2 estimation. Its structure is similar to our Model 3 but with fewer layers. Table 1 reveals that Model 2 is the best in terms of correlation in both PD and PU cases, whereas Model 3 achieves the best in MAE and RMSE, suggesting that Model 2 and Model 3 are comparably the best in the individualized learning. All of our model configurations outperform Ding et al. [10]. For example, in the PU case for Model 3, the correlation is improved from 0.34 to 0.41 and the MAE is lowered from 3.40% to 1.81%. It is worth noting that the international standard for clinically acceptable pulse oximeters tolerates an error of 4% [13], and our estimation errors are all within this range.

There are two factors, including the skin type and the side of the hand, that might influence the performance of SpO_2 estimation. We therefore analyze the following two questions: (1) Whether the different skin types matter in PU or PD case, and (2) whether the side of hand matters in lighter skin (types II + III) or darker skin (types IV + V). The box plots in Fig. 7 shows the distributions of the test correlations from all the three proposed models in PU and PD modes of (a) lighter skin and darker skin participants, and (b) all participants.

To answer question (1), we focus on the left panel of Fig. 7a. We note that overall, the medians of darker skin group are larger than those of the lighter skin group. Zooming into the PD case, we can confirm that the darker skin group indeed outperforms the light group since the former has a smaller interquartile range (IQR). However, for the PU case, the no significant performance difference can be observed, because while the dark skin group is better in a larger median, the light skin group is better in a narrower IQR. To answer question (2), we first focus on the left panel of Fig. 7b. We note that no significant performance difference can be observed between PD and PU given one has a better median and the other has a better IQR, when participants of all skin colors are considered together. However, if we zoom into the subset of darker skin group as shown in the left panel of Fig. 7a, we observe that PD is better than PU given its higher median and narrower IQR. To summarize, in the participant-specific experiments, (1) darker skin group outperforms the lighter skin group when using the back side of the hand as the ROI for SpO_2 prediction but they are comparable when using the palm of the hand; and (2) the side of the hand has an impact on SpO_2 prediction for the darker skin group but not for the lighter skin group.

Table 2: Performance comparison of each model structure in leave-one-participant-out experiments. Results are given as the median of all participants.

	Hand Mode	Correlation	MAE (%)	RMSE (%)
Model 1 (Proposed)	PD	**0.33**	**2.33**	3.07
	PU	**0.46**	**1.97**	**2.16**
Model 2 (Proposed)	PD	0.15	2.43	3.35
	PU	0.33	2.08	2.41
Model 3 (Proposed)	PD	0.23	2.48	**2.98**
	PU	0.27	2.02	2.54
Ding et al.	PD	0.11	3.19	3.76
	PU	0.26	2.43	2.85

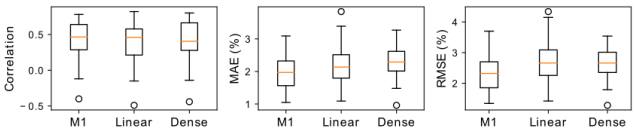


Figure 8: Ablation studies results. Comparison between the proposed (nonlinear) Model 1 (M1), modified M1 with only linear channel combinations, and modified M1 with fully connected dense layers instead of convolutional layers. Ablation studies confirm that the nonlinear channel combinations and convolutional layers improve model performance.

4.3. Leave-One-Participant-Out Results

To investigate whether the features learned by the model from other participants are generalizable to new participants whom it has not seen before, we conduct leave-one-participant-out experiments. For each experiment, when testing on a certain participant, we use all the other participant's data for training and leave the test participant's data out. The recordings from all the non-test participants are used for participant-wise cross-validation to select the best model structure and hyperparameters. The selected model is evaluated on the two recordings of the test participant, whose data was never seen by the model during training.

Table 2 shows the performance comparison of each model in leave-one-participant-out experiments. Model 1 achieved the best performance in terms of correlation and MAE, and achieved the best RMSE for the PU case. Model 3 achieved better RMSE results for the PD case but the correlation result was low, suggesting that the model achieved low error by simply predicting a constant SpO_2 near the middle of the SpO_2 range. The best performance of Model 1 in the leave-one-participant-out experiment may imply that the features extracted after combining the color channels at the beginning of the pipeline can be generalized better to unseen participants than the features extracted before channel combination or through interleaving as in Models 2 or 3. In the participant-specific case, the model is specifically tailored to the test individual, whereas the leave-one-participant-out case is more difficult because the the model needs to accommodate for the variation in the population. As expected, in Fig. 7, we observe that the overall results from the leave-one-participant-out experiments do not match those from the participant-specific experiments. Because of the modest size of the dataset, the model has not seen as diverse data as a larger and richer dataset would offer. The generalization capability to new participants can be improved when more data is available.

We now revisit the two research questions raised in Section 4.2 under the leave-one-participant-out setup. First, we analyze the impact of skin type given the same side of the hand. From the right panel of Fig. 7a, we observe that in the PD case, the darker skin group outperforms the lighter skin group , whereas in PU case, the performances are comparable. This observation is consistent with the participant-specific experiments that when using the palm as the ROI, the skin color is not a factor to the accuracy of SpO_2 estimation. Second, we analyze the impact of the side of the hand for two skin color groups. The right panel of Fig. 7a reveals that for darker skin group, the PD case outperforms the PU case, which is consistent with the results from the participant-specific experiments. However, in contrast to these experiments, the PU outperforms the PD in both lighter skin group as well as the mixed group as illustrated in the right panel of Fig. 7b. This different generalization capability in the PU and PD cases may be attributed to skin color difference between the palm and the back of the hand. The color of the back of the hand tends to be darker than the color of the palms, and has larger color variation among participants due to different degrees of sunlight exposure. In contrast, the color variation of the palms is much milder among participants. Furthermore, in the participant-specific experiments, the individualized models learn the traits of the skin type and the side of the hand from each participant, whereas in the leave-one-participant-out experiments, the learned model must capture the general characteristics of the population.

4.4. Ablation Studies

To justify the use of nonlinear channel combinations and convolutional layers for temporal feature extraction in our proposed models, we conduct two ablation studies comparing the performance of these model components to other generic ones. We focus on the PU case to avoid uncontrolled impact of such factors as the skin tone and hair. In the first ablation study, we compare nonlinear to linear channel combination. We create a variant of Model 1 with only a single linear channel combination layer with no activation function and repeat the leave-one-participant-out experiments. Fig. 8 reveals that our proposed Model 1 with nonlinear channel combination layers outperforms the variant of Model 1 in terms of correlation, MAE, and RMSE. In the second study, we compare the performance of using convolutional layers for temporal feature extraction to using fully-connected dense layers. We create another variant of Model 1 and repeat leave-one-participant-out experiments.

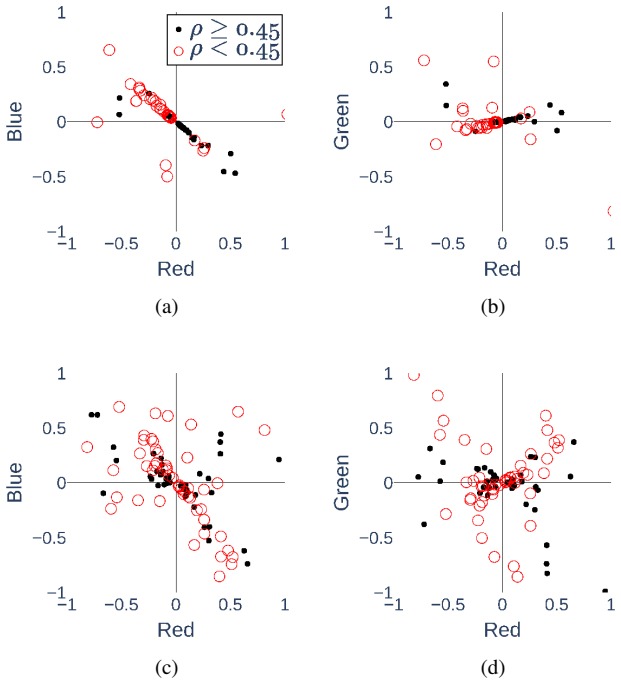


Figure 9: Learned RGB channel weights. Plots (a) and (b) are the channel weights learned by different model instances trained on the data of all study participants together, projected onto the RB and RG planes in the RGB space. Plots (c) and (d) are the RB and RG projections of the learned channel weights for model instances trained on random subsets of the participants' data. Each point is color coded according to the correlation ρ achieved by the instance.

Fig. 8 reveals that our proposed Model 1 with convolutional layers achieve better performance across all three metrics. A summary of the studies results can be found in the supplemental document.

5. Visualizations of RGB Combination Weights

To understand and explain what our physiologically inspired models have learned, we conduct a separate investigation to visualize the learned weights for the RGB channels. Our goal is to understand the best way to combine the RGB channels for SpO_2 prediction. Having an explainable model is important for a physiological prediction task like this. Our neural network models can be considered as nonlinear approximations of the hypothetically true function that can extract the physiological features related to SpO_2 buried in the RGB videos. The ratio-of-ratios method, for example, is another such extractor that combines the information from the different color channels at the end of the pipeline. For this experiment, we use the modified version of Model 1 from the ablation studies that has only a single linear channel combination at the beginning. Seeing that using a single linear channel combination did not significantly reduce model performance in the ablation studies, and understanding that the linear component may dominate the Taylor expansion of a nonlinear function, we use only linear combinations for this model to facilitate more interpretable visualizations.

We have trained 100 different instances of the model on the first two cycles from all the recordings and tested on the third cycle from all recordings. The difference between each instance is that the weights are randomly initialized. The weights for each channel learned by the model instances were visualized as points representing the heads of the linear combination vector in RGB space. Each point is colored according to the average test correlation achieved by the model instance. Figs. 9a and 9b show the projections of these points onto the RB and RG planes. The subfigures reveal that the majority of the channel weights lay along certain line in the RGB space. For the weights on the line the ratio of the blue channel weight to the red channel weight is 0.87, the ratio of the green channel weight to red channel weight is 0.18. It is clear that the red and blue channels are the dominating factors for SpO_2 prediction.

To further verify this result, we repeat this experiment but instead of using the data from all participants, for each model instance we randomly select seven participants and use their data for training and testing. In this case, the difference between each model instance is not only the initialized weights but also the random subset of participants that the model was trained on. Fig. 9d reveals that most of the better performing instances (with $\rho \geq 0.45$) have little contribution from the green channel. In Fig. 9c, we again see that most of the points lay on a line in the RB plane, the ratio of the blue channel weight to the red channel weight for these points is 0.80.

These results are in accordance with the physical understanding of how light is absorbed by hemoglobin in the blood. Fig. 2 reveals a large difference between the extinction coefficients, or the amount of light absorbed, by deoxygenated and oxygenated hemoglobin at the red wavelength. There is a significantly smaller difference at the blue wavelength and almost no difference at green. The amount of light absorbed influences the amount of light reflected which can be measured through the camera. A larger difference in extinction coefficients makes it easier to measure the ratio of light absorbed by oxygenated vs. deoxygenated hemoglobin over time. This ratio indicates the level of blood oxygen saturation. Therefore, from a physiological perspective, it makes sense for the neural networks to give larger weight to the red and then blue channels and give little to the green channel. These visualizations indicate that the models are learning physically meaningful features.

6. Conclusion

In this paper, we have proposed the first CNN-based work to solve the challenging problem of video-based re-

mote SpO_2 estimation. We have designed three optophysiologically inspired neural network architectures. In both participant-specific and leave-one-participant-out experiments, our models are able to achieve better results than the state-of-the-art method. We have also analyzed the effect of skin color and the side of the hand on SpO_2 estimation and have found that in the leave-one-participant-out experiments, the side of the hand plays an important role with better SpO_2 estimation results achieved in the PU case. We have also shown the explainability of our designed architectures by visualizing the weights for the RGB channel combinations learned by the neural network, and have confirmed that the choice of the color band learned by the neural network is consistent with the established optophysiological methods.

References

[1] Optical Absorption of Hemoglobin. https://omlc.org/spectra/hemoglobin/. Accessed: 2021-03-09. 3

[2] Australian Radiation Protection and Nuclear Safety Agency. *Fitzpatrick skin phototype*. 5, 12

[3] Ufuk Bal. Non-contact estimation of heart rate and oxygen saturation using ambient light. *Biomedical Optics Express*, Jan. 2015. 1, 2

[4] Wilhelm Burger and Mark J. Burge. *Digital Image Processing - An Algorithmic Introduction using Java.* Springer, 2008. 3

[5] Gabriella Casalino, Giovanna Castellano, and Gianluca Zaza. A mHealth solution for contact-less self-monitoring of blood oxygen saturation. In *IEEE Symposium on Computers and Communications (ISCC)*, Jul. 2020. 1

[6] Mingliang Chen, Qiang Zhu, Min Wu, and Quanzeng Wang. Modulation model of the photoplethysmography signal for vital sign extraction. *IEEE Journal of Biomedical and Health Informatics*, Aug. 2020. 1

[7] Weixuan Chen and Daniel McDuff. DeepPhys: Video-based physiological measurement using convolutional attention networks. In *The European Conference on Computer Vision (ECCV)*, pages 349–365, 2018. 2

[8] Jennifer Couzin-Frankel. The Mystery of The Pandemic's "Happy Hypoxia". *Science*, 2020. 1

[9] Gerard De Haan and Vincent Jeanne. Robust pulse rate from chrominance-based rppg. *IEEE Transactions on Biomedical Engineering*, Jun. 2013. 1

[10] Xinyi Ding, Damoun Nassehi, and Eric C Larson. Measuring oxygen saturation with smartphone cameras using convolutional neural networks. *IEEE Journal of Biomedical and Health Informatics*, Dec. 2018. 1, 2, 3, 6

[11] Riccardo Favilla, Veronica Chiara Zuccala, and Giuseppe Coppini. Heart rate and heart rate variability from single-channel video and ica integration of multiple signals. *IEEE journal of biomedical and health informatics*, Nov. 2018. 1

[12] B.W. Green, P. J.; Silverman. *Nonparametric Regression and Generalized Linear Models*. Chapman and Hall, 1990. 5

[13] International Organization for Standardization. *Particular requirements for basic safety and essential performance of pulse oximeter equipment* , 2011. 6

[14] Luca Iozzia, Luca Cerina, and Luca Mainardi. Relationships between heart-rate variability and pulse-rate variability obtained from video-ppg signal using zca. *Physiological measurement*, Sep. 2016. 1

[15] Lingqin Kong, Yuejin Zhao, Liquan Dong, Yiyun Jian, Xiaoli Jin, Bing Li, Yun Feng, Ming Liu, Xiaohua Liu, and Hong Wu. Non-contact detection of oxygen saturation based on visible light imaging device using ambient light. *Opt. Exp.*, Jul. 2013. 1

[16] Lisha Li, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. Hyperband: A novel bandit-based approach to hyperparameter optimization. *The Journal of Machine Learning Research*, Apr. 2018. 4

[17] Xiaobai Li, Jie Chen, Guoying Zhao, and Matti Pietikainen. Remote heart rate measurement from face videos under realistic situations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4264–4271, 2014. 1

[18] Zhiyuan Lu, Xiang Chen, Zhongfei Dong, Zhangyan Zhao, and Xu Zhang. A prototype of reflection pulse oximeter designed for mobile healthcare. *IEEE Journal of Biomedical and Health Informatics*, Aug. 2015. 1, 2

[19] Yunyoung Nam, Bersain A Reyes, and Ki H Chon. Estimation of respiratory rates using the built-in microphone of a smartphone or headset. *IEEE Journal of Biomedical and Health Informatics*, Sep. 2015. 3

[20] Meir Nitzan, Ayal Romem, and Robert Koppel. Pulse oximetry: Fundamentals and technology update. *Medical Devices (Auckland, NZ)*, 7:231, 2014. 1, 2

[21] Xuesong Niu, Shiguang Shan, Hu Han, and Xilin Chen. RhythmNet: End-to-end heart rate estimation from face via spatial-temporal representation. *IEEE Trans. on Image Processing*, Oct. 2019. 2, 3

[22] Nobuyuki Otsu. A threshold Selection Method from Gray-level Histograms. *IEEE Trans. Syst., Man, and Cybernet.*, Jan. 1979. 3

[23] Ming-Zher Poh, Daniel J McDuff, and Rosalind W Picard. Advancements in noncontact, multiparameter physiological measurements using a webcam. *IEEE transactions on biomedical engineering*, Oct. 2010. 1

[24] Christopher G Scully, Jinseok Lee, Joseph Meyer, Alexander M Gorbach, Domhnull Granquist-Fraser, Yitzhak Mendelson, and Ki H Chon. Physiological parameter monitoring from optical recordings with a mobile phone. *IEEE Trans. on Biomedical Eng.*, Jul. 2011. 1, 2

[25] John W Severinghaus. Takuo Aoyagi: Discovery of pulse oximetry. *Anesthesia & Analgesia*, Dec. 2007. 1, 2, 3

[26] Dangdang Shao, Chenbin Liu, Francis Tsow, Yuting Yang, Zijian Du, Rafael Iriya, Hui Yu, and Nongjian Tao. Noncontact Monitoring of Blood Oxygen Saturation Using Camera and Dual-wavelength Imaging System. *IEEE Trans. Biomed. Eng*, Sep. 2015. 1

[27] Kwanghyun Sohn, Faisal M Merchant, Omid Sayadi, Dheeraj Puppala, Rajiv Doddamani, Ashish Sahani, Jagmeet P Singh, E Kevin Heist, Eric M Isselbacher, and Antonis A Armoundas. A novel point-of-care smartphone based system for monitoring the cardiac and respiratory systems. *Scientific Reports*, Mar. 2017. 3

[28] Radim Špetlík, Vojtech Franc, and Jirí Matas. Visual heart rate estimation with convolutional neural network. In *British Machine Vision Conf., Newcastle, UK*, Sep. 2018. 2

[29] Nichole Starr, Daniela Rebollo, Yohannes Molla Asemu, Leulayehu Akalu, Hanan Ali Mohammed, Misrak Woldeyohannes Menchamo, Eyayelem Melese, Senait Bitew, Iain Wilson, Mahelet Tadesse, et al. Pulse oximetry in low-resource settings during the COVID-19 pandemic. *The Lancet Global Health*, Sep. 2020. 1

[30] Lionel Tarassenko, Mauricio Villarroel, Alessandro Guazzi, João Jorge, DA Clifton, and Chris Pugh. Non-contact Video-based Vital Sign Monitoring Using Ambient Light and Auto-regressive Models. *Physiol. Meas*, Mar. 2014. 1, 2

[31] Hsin-Yi Tsai, Kuo-Cheng Huang, and J Andrew Yeh. No-contact oxygen saturation measuring technology for skin tissue and its application. *IEEE Instrum. Meas. Magazine*, Sep. 2016. 1

[32] Sergey Tulyakov, Xavier Alameda-Pineda, Elisa Ricci, Lijun Yin, Jeffrey F Cohn, and Nicu Sebe. Self-adaptive matrix completion for heart rate estimation from face videos under realistic conditions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2396–2404, 2016. 1

[33] Mark Van Gastel, Sander Stuijk, and Gerard De Haan. New Principle for Measuring Arterial Blood Oxygenation, Enabling Motion-Robust Remote Monitoring. *Scientific Reports*, Dec. 2016. 1, 2

[34] Mark van Gastel, Wim Verkruysse, and Gerard de Haan. Data-driven Calibration Estimation for Robust Remote Pulse-oximetry. *Applied Sciences*, Jan. 2019. 1

[35] Wenjin Wang, Albertus C den Brinker, Sander Stuijk, and Gerard De Haan. Algorithmic principles of remote PPG. *IEEE Trans. on Biomedical Eng.*, Sep. 2016. 3

[36] John G Webster. *Design of Pulse Oximeters*. CRC Press, Oct. 1997. 2, 3

[37] Qiang Zhu, Chau-Wai Wong, Chang-Hong Fu, and Min Wu. Fitness heart rate measurement using face videos. In *IEEE Int'l Conf. on Image Proc. (ICIP)*, Sep. 2017. 1, 3

7. Supplemental Material

7.1. Fitzpatrick Skin Types

Our self-collected dataset consists of hand video recordings and SpO_2 data from fourteen participants, of which there were six males and eight females between the ages of 21 and 30. Participants were asked to categorize their skin tone based on the Fitzpatrick skin types [2] shown in Fig. 10. The Fitzpatrick skin types classify the skin by its reaction to exposure to sunlight and pigmentation. From type I to type VI, the skin color becomes darker and less prone to be burned by the sunlight. Among the fourteen participants, two are from type II, eight are from type III, one is from type IV, and three are from type V.

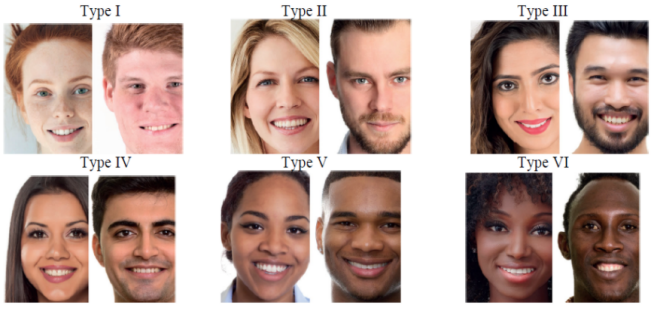


Figure 10: Fitzpatrick skin types. Reproduced from [2].

7.2. Skin Type and Palm Mode Comparison

Table 3 is a supplement to Fig. 7 in the main paper with numerical values specified for the factor analysis, including the skin type and the side of the hand. This table presents the comparison of the test correlations from all the three proposed models in palm up (PU) and palm down (PD) data collection modes of lighter skin participants (type II and III), darker skin participants (type IV and V), and all participants. The results from participant-specific and leave-one-participant-out experiments are presented in the median and interquartile range (IQR). IQR measures the difference between the first quartile and the third quartile, and quantifies the spread of the distribution. We analyze the following two questions from the results in Table 3: (i) Whether the different skin types matter in PU or PD case, and (ii) whether the side of hand matters in lighter skin or darker skin.

Note that the following analysis based on Table 3 is the quantitative version of the last paragraphs in Section 4.2 and 4.3 of the main paper. Conclusions are exactly the same.

To answer the first question, we first focus on the top panel of Table 3 to examine the participant-specific case. In the PD case, the darker skin group outperforms the light group since the former has a larger median of 0.48 and a smaller IQR of 0.20. In the PU case, the medians of the lighter skin group and darker skin group are 0.41 and 0.45, with IQR being 0.30 and 0.38, respectively. Even though

Table 3: Comparison of correlations for lighter vs. darker skin types vs. all skin types in both PU and PD cases. This table is a supplement to Figure 7 of the main paper.

	Participant-Specific					
	Lighter		Darker		Overall	
Hand Mode	Median	IQR	Median	IQR	Median	IQR
PD	0.44	0.50	0.48	0.20	0.45	0.41
PU	0.41	0.30	0.45	0.38	0.41	0.33
	Leave-One-Participant-Out					
	Lighter		Darker		Overall	
Hand Mode	Median	IQR	Median	IQR	Median	IQR
PD	0.14	0.41	0.43	0.19	0.24	0.39
PU	0.35	0.46	0.31	0.31	0.34	0.42

the median from the darker group is 9.8% higher, the IQR is 26.7% worse. Thus, no significant performance difference is observed in the PU case. Next, we focus on the bottom panel of Table 3 to analyze the results from the leave-one-participant-out experiment. We observe that in the PD case, the darker skin group with a median of 0.43 outperforms the lighter skin group with a median of 0.14, whereas in the PU case, the performances are comparable. This observation is consistent with the participant-specific experiments that when using the palm as the ROI, the skin color is not a factor to the accuracy of SpO_2 estimation.

To answer the second question, we first focus on the participant-specific case in the top panel of Table 3. In both lighter skin group and the mixed group, the medians of PU cases, which are 0.41 and 0.41, respectively, are smaller than PD cases with medians of 0.44 and 0.45, while the IQRs are also narrower (PU: 0.30 and 0.33 vs. PD: 0.50 and 0.41), which make the distributions comparable. In the darker skin group, the medians of PU and PD cases are 0.45 and 0.48, with IQR being 0.38 and 0.20, respectively. In comparison, we find that there is no significant difference between PU and PD cases in our current lighter skin and overall groups, whereas in the darker skin group, the PD case is better than the PU case. Second, we focus on the results under the leave-one-participant-out setup in the bottom panel of Table 3. For the darker skin group, the PD case with a median of 0.43 and IQR of 0.19 outperforms the PU case with a median of 0.31 and IQR of 0.31, which is consistent with the results from the participant-specific experiments. In contrast, in both the lighter skin group and the mixed group, the PU cases with medians of 0.35 and 0.34 significantly outperform the PD cases with medians of 0.14 and 0.24.

7.3. Ablation Studies

Table 4 supplements Fig. 8 of the main paper for the ablation studies of Model 1 in the leave-one-participant-out setup. This table presents the medians and IQRs specified for numerical comparison. The ablation studies justify the

use of (i) nonlinear channel combinations and (ii) convolutional layers for temporal feature extraction. In ablation study 1, we replace the nonlinear channel combination with a single linear channel combination layer with no activation function as the first variant of Model 1. In ablation study 2, we replace the convolutional layers for temporal feature extraction with fully-connected dense layers as the second variant of Model 1.

Table 4: Ablation studies for Model 1 in the leave-one-participant-out mode. This table is a supplement to Figure 8 of the main paper.

Method		ρ	MAE(%)	RMSE(%)
Linear Ch. Comb.	Median	0.46	2.14	2.66
+ Conv. layer for Feat. Extra.	IQR	0.38	0.73	0.93
Nonlinear Ch. Comb.	Median	0.41	2.29	2.66
+ Fully Connec. layer for Feat. Extra.	IQR	0.39	0.63	0.70
Model 1 (Proposed): Nonlinear Ch. Comb.	Median	**0.46**	**1.97**	**2.32**
+ Conv. layer for Feat. Extra.	IQR	0.36	0.80	0.87

Note that the following analysis based on Table 4 is the quantitative version of the first paragraph in Section 4.4 of the main paper. Conclusions are exactly the same.

First, we compare the first and the third rows in Table 4 for ablation study 1. Our proposed Model 1 achieves a better correlation with a median of 0.46 and IQR of 0.36 and a better RMSE with a median of 2.32 and IQR of 0.87 than its linear channel combination variant. Besides, Model 1 achieves a comparable MAE with a better median of 1.97 but a wider IQR of 0.80. The overall better performance of Model 1 suggests the necessity of using the nonlinear channel combination method. Second, in ablation study 2, we compare the second and the third rows in Table 4. We observe that Model 1 outperforms its second variant with fully-connected layers for feature extraction with better medians in terms of correlation (0.46 vs. 0.41), MAE (1.97 vs. 2.29), and RMSE (2.32 vs. 2.66) and narrower IQR of correlation. This suggests that convolutional layers are better than fully connected layers for temporal feature extraction.