Chapter 5

Optophysiological Model Guided Neural Networks for Contactless Blood Oxygen Estimation From Hand Videos

5.1 Introduction

Deep learning has demonstrated promising performance in camera-based physiological measurements, such as heart rate, breathing rate, and body temperature [26, 101, 127, 162]. An end-to-end convolutional attention network was proposed in [26] 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 [101] demonstrates that the heart rate can be directly inferred using a convolutional network with spatial-temporal representations of the face videos as its input. Mobile applications have been developed to utilize CNNs to measure body temperature from facial images [162].

Deep learning for SpO_2 monitoring from videos is still in its early stage. Ding *et al.* [37] proposed a convolutional neural network architecture for contact-based SpO_2 monitoring with smartphone cameras. Even though the work in [37] showed better per-

formance than the conventional ratio-of-ratios method, their technique requires the users' fingertips to be in contact with the illuminated flashlight and camera, which not only may lead to a sense of burning for a continuous period of time but also raises sanitation concerns, especially if the sensing device is shared by multiple participants during pandemics.

The video-based contactless methods for physiological signal sensing provide a comfortable and an unobtrusive option of monitoring SpO_2 and have the potential to be adopted in health screening and telehealth. In Chapter 4, we have taken advantage of the contact-free sensing from a regular RGB camera as well as the well-known two-channel RoR mechanism from pulse oximeters for accurate SpO_2 estimation. In particular, we proposed a strategic use of the hand video data by performing spatial and temporal data analysis of more than two color channels. Under the umbrella of the synergistic framework that takes advantage of both biophysical imaging principles and the availability of participants' video and SpO_2 data to learn and determine the details for obtaining SpO_2-relevant features and making SpO_2 estimation, Chapter 4 determined the specific features and the related detailed parameters **explicitly** from the biophysical imaging principles, while in this chapter, we propose to use these principles to guide the design of neural network architectures to "learn" the specific SpO_2-relevant features from the input video signals with a data-driven **implicit** approach and perform SpO_2 estimation in a holistic manner. Compared to the principled signal processing scheme for feature engineering proposed in Chapter 4, the neural network based schemes proposed in this chapter learn features implicitly from data and use synergy with the principled methodology to guide the selection of the neural network architectures.

Specifically, inspired by the optophysiological model for SpO_2 measurement [117, 142, 152], in this chapter, we develop convolutional neural networks (CNN) based SpO_2 estimation schemes designed based on the optophysiological models for a better explanation, wherein data-driven feature extraction and estimation of the blood oxygen saturation level comprise implementing a combination of spatial averaging, color channel mixing, and temporal trend analysis. The schemes analyze the videos of a participant's hand captured by regular RGB cameras in a contactless way, which is convenient and comfortable for users and can protect their privacy and allow for keeping face masks on.

5.2 Proposed Optophysiology-Guided Neural Network Method for estimating SpO_2 From Videos

Fig. 5.1 is an overview of the system design. First, the ROI, including the palm and the back of the hand, is extracted from the smartphone captured videos. Second, the ROI is spatially averaged to produce R, G, and B color time series. Next, the three color-channel signals are fed into an optophysiology-inspired CNN to extract features to achieve more explainable and accurate SpO_2 predictions.

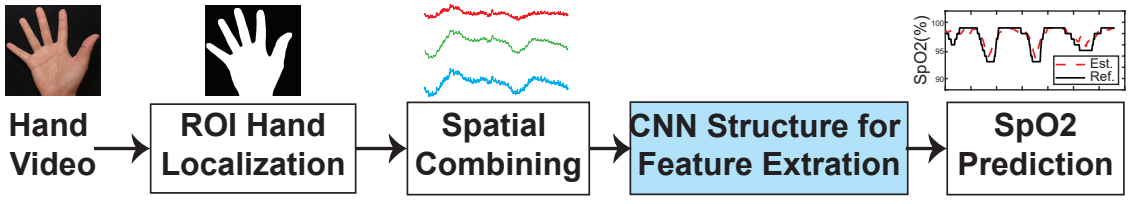


Figure 5.1: Proposed neural network based contactless SpO_2 estimation method. Three color time series are extracted from the skin area of a hand video by spatial averaging and are then fed into an optophysiology-inspired neural network to extract features by color channel mixing and temporal analysis for SpO_2 prediction.

5.2.1 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*. As Chapter 4.3 explained, the ROI of the skin pixels is separated using Otsu's method [103] which determines a threshold that best separates the skin pixels from the background by minimizing the variance within the skin and non-skin classes in the Cr axis of the YCbCr color space [18]. 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.

In this chapter, 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 an 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 [121]. 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.

5.2.2 Neural Network Architectures

The previous neural network work for SpO_2 prediction mainly explored prediction, but not the model explainability [37]. 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 that 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 [101, 166] and respiratory rate sensing [96, 126], the RGB skin color signals are often combined first to form one "rPPG" signal followed by temporal feature extraction, as is done in the plane-orthogonal-to-skin (POS) algorithm [149]. In contrast, for conventional SpO_2 sensing methods such as the ratio-of-ratios [152], the temporal features are extracted first (i.e., extracting AC and DC from a time window for each color channel) and the color components are combined at the end (i.e., taking the ratio and pairwise ratio of ratios) before doing regression fitting. 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.

Color Channel Mixing Followed by Temporal Analysis: In Model 1, shown as the leftmost structure depicted in Fig. 5.2, we combine the color channels first using several channel combination layers and then extract temporal features using temporal convolution

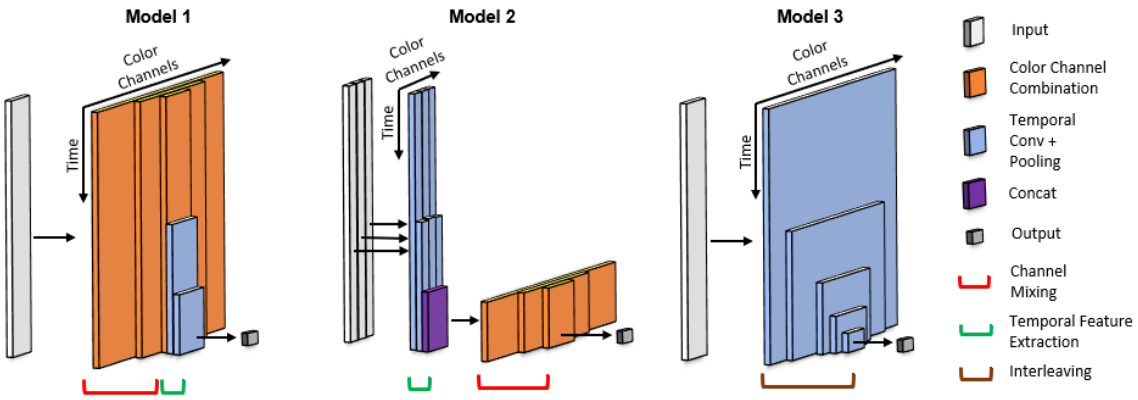


Figure 5.2: Proposed network structures for predicting SpO_2 levels from a fixed-length segment of skin color signals. We highlight the differences among the three model configurations instead of showing the exact model structures. Model 1 combines the RGB channels before temporal feature extraction. 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.

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:

$$\mathbf{V} = \sigma(\mathbf{W}\mathbf{U} + \mathbf{b}\mathbb{1}^T), \tag{5.1}$$

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 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 point in the 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, 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. This model has three channel combination layers, three feature extraction layers, and a total of 34K trainable parameters.

Temporal Analysis Followed by Color Channel Mixing: In Model 2, which is the middle structure depicted in Fig. 5.2, we reverse the order of color 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. This model has three channel combination layers, 2 feature extraction layers, and a total of 12K parameters.

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. 5.2, 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. This model has 4 layers and a total of 307K parameters.

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 that has 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 [86], which allows for a faster and more efficient search over a large parameter space than grid search or random search. It does this by running random parameter configurations on a specific schedule of iterations per configuration and uses earlier results to select candidates for longer runs. The parameters that are 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.

5.3 Experimental Results

5.3.1 Dataset and Capturing Conditions

Our proposed models are evaluated on a self-collected dataset that is studied in Chapter 4. To recapitulate, the dataset consisted of two sessions of hand video record-

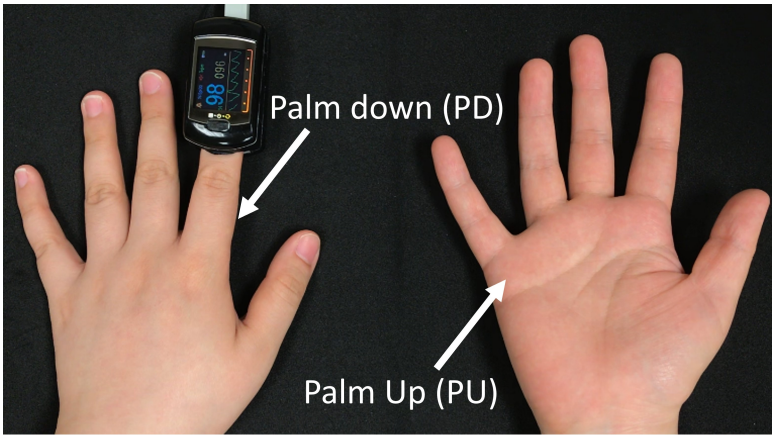
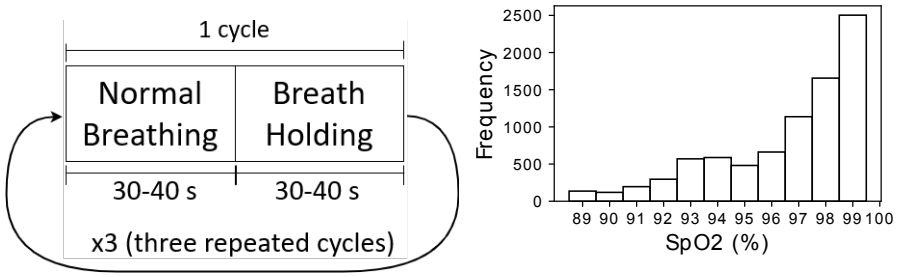


Figure 5.3: Illustration of two hand-video capturing positions. Left hand: palm down (PD). Right hand: palm up (PU).

ings and simultaneously recorded reference SpO_2 data from each of the 14 participants, of which there were six males and eight females between the ages of 21 and 30. 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 using protocol #1376735 approved by the University of Maryland Institutional Review Board (IRB).

Each participant was asked to place his/her hands still on a table to avoid hand motion. Their palm of the right hand and the back of the left hand are facing the camera, as illustrated in Fig. 5.3. 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. 5.4(a). The participant breathes normally for 30–40 seconds, exhales all the way, and then holds his/her breath for 30–40 seconds. This process is repeated three times for each session. The collected SpO_2 value distribution is shown in Fig. 5.4(b).

In this chapter, we increase the data size by interpolating the reference SpO_2 signal



(a) (b)

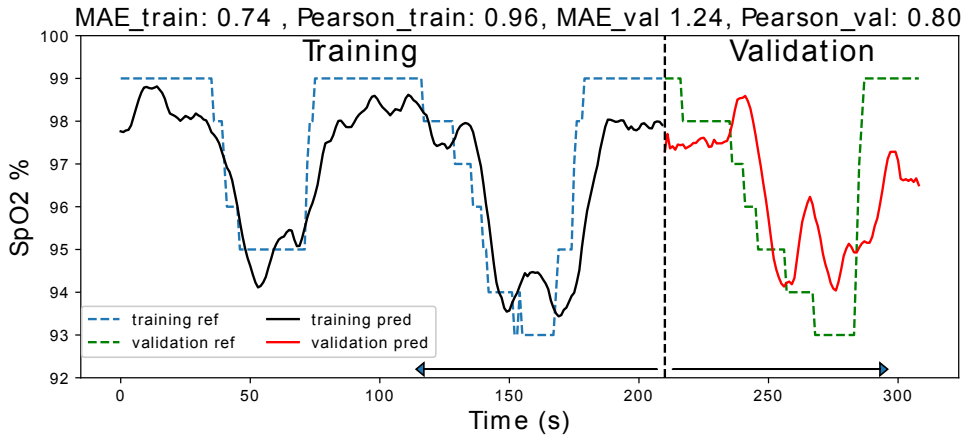
Figure 5.4: (a) Breathing protocol that participants were asked to follow, including 3 cycles of normal breathing and breath holding. (b) Histogram of SpO_2 values in the collected dataset.

to 5 sample points per second to match the segment sampling rate (Chapter 5.2.1) using a smooth spline approximation [50]. 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 video recording, the model is sequentially fed with 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% based on physiological knowledge. A 10-second long moving average filter is applied to generate a refined time series of predicted SpO_2 values.

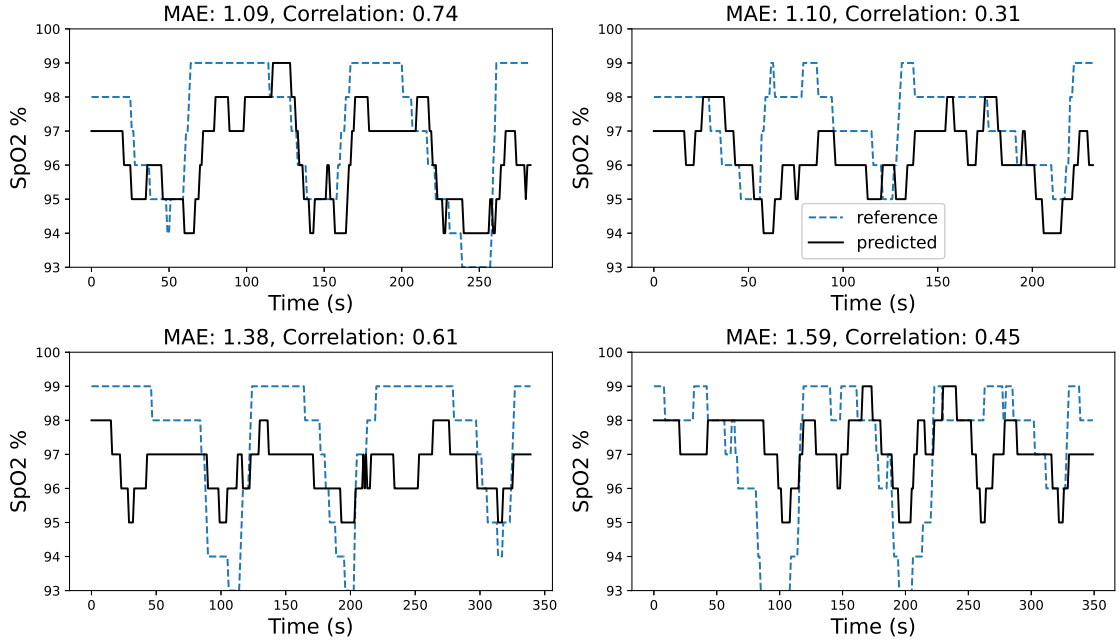
5.3.2 Participant-Specific Results

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 least 15 minutes in between. One recording is used for training and validation of the model and



(a)



(b)

Figure 5.5: (a) Training vs. validation predictions. (b) Test predictions of varying performance with reference SpO_2. The higher the Pearson correlation, the better the prediction captures the reference SpO_2 trend. The lower the MAE, the better the prediction captures the dips in SpO_2.

the remaining recording is for testing. An example of the training and validation predictions curves is shown in Fig. 5.5(a). 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. 5.5(b). 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. This is an example of the bootstrapping data reuse strategy [67, Chapter 5]. The oversampling also helps address the imbalance in SpO_2 data values that is shown in Fig. 5.4(b).

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, and 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 5.1 shows the performance comparison of our proposed models with the prior-art model from Ding *et al.* [37]. To the best of our knowledge, Ding *et al.*'s model is the only convolutional neural network structure that has been tried for contact-based SpO_2 estimation. Its structure is similar to our Model 3 but with fewer layers. We also compare with the classic ratio-of-ratios method proposed by Scully *et al.* [117]. The performance is measured in Pearson's Correlation, mean absolute error (MAE), and root mean square

	Hand Mode	Correlation		MAE (%)		RMSE (%)	
		Median	IQR	Median	IQR	Median	IQR
Model 1 (Proposed)	PD	0.41	0.40	2.12	0.91	2.51	0.78
	PU	0.39	0.37	2.16	1.80	2.70	2.09
Model 2 (Proposed)	PD	**0.46**	0.44	2.09	1.32	2.52	1.63
	PU	**0.41**	0.32	1.96	0.68	2.48	0.89
Model 3 (Proposed)	PD	0.44	0.40	**1.93**	1.11	2.48	1.31
	PU	**0.41**	0.46	**1.81**	1.83	2.43	2.44
Scully *et al.* [117]	PD	0.08	0.37	1.94	0.92	**2.22**	0.77
	PU	0.19	0.24	2.01	0.80	**2.36**	0.78
Ding *et al.* [37]	PD	0.38	0.39	3.25	2.85	3.83	3.24
	PU	0.34	0.56	3.40	3.16	4.58	3.12

Table 5.1: Performance comparison of each model structure for participant-specific experiments. Results are given as the test median and IQR of all participants.

error (RMSE), and the results of each condition are summarized in the median and interquartile range (IQR). IQR quantifies the spread of an empirical distribution of a set of data points by computing the difference between the first quartile and the third quartile of the distribution.

Table 5.1 reveals that Model 2 achieves the best correlation in both PD and PU cases, whereas Model 3 achieves the best MAE and a comparable correlation with Model 2, suggesting that Model 2 and Model 3 are comparably the best in the individualized learning. Even though the method proposed in Scully *et al.* [117] achieves the best (lowest) RMSE, its correlations are the worst (lowest). This suggests that the classic ratio-of-ratios method cannot track the trend of SpO_2 well using the contactless measurement by smartphone. All of our model configurations outperform Ding *et al.* [37]. 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 allows an error of 4% [65], and our estimation errors are all within this range.

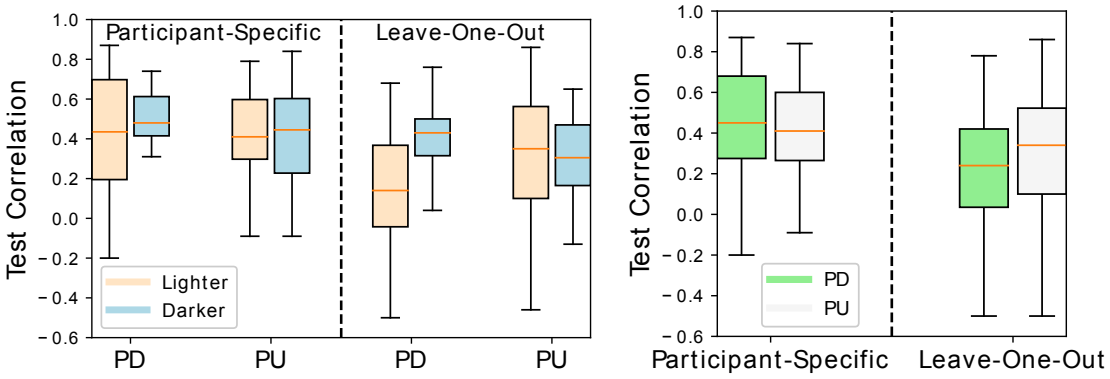


Figure 5.6: Boxplots comparing distributions of correlations for (a) lighter vs. darker skin types, and (b) PD vs. PU for all skin types. The PD results are better for darker skin tones in both the participant-specific and leave-one-out cases.

There are two factors, including the skin type and the side of the hand, which might influence the performance of SpO_2 estimation. We therefore investigate the following two questions: (1) Whether the different skin types matter in PU or PD cases, 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. 5.6 summarize 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.

Bayesian statistical test: We use Bayesian statistical tests to further analyze the results in Fig. 5.6 by providing a probabilistic assessment of whether the results from two groups being compared have the same mean [78–82]. We avoid using the popular t-test because it makes only a binary decision due to its lack of direct information about the probability of difference between group means of the given data [78]. In contrast, the Bayesian statistical test computes the posterior distribution of difference between the two group means to quantify its certainty of possible values [82]. The decision rule of the Bayesian

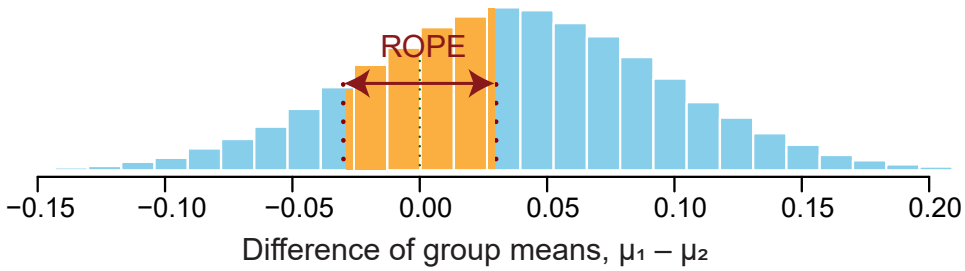


Figure 5.7: Posterior distribution of the difference of group means. This shows an example of an undecided case of the Bayesian statistical test given that the ROPE of zero difference is set to $[-0.03, 0.03]$ and 33% of the posterior distribution falls within the ROPE. The percentage of coverage can be used to quantify the certainty that two groups have the same mean.

statistical test for the null hypothesis that the two groups have the same mean can be stated as follows given the *region of practical equivalence (ROPE)* of zero difference [80]:

- (*Accepted*): If the percentage of the posterior distribution of the group-mean difference inside the ROPE is sufficiently high (e.g., greater than 95% [80]), then the null hypothesis is accepted.

- (*Rejected*): If the percentage of the posterior distribution of the group-mean difference inside the ROPE is sufficiently low (e.g., less than 2.5% [80]), then the null hypothesis is rejected.

- (*Undecided*): When the null hypothesis is neither accepted nor rejected, the percentage of the posterior distribution of the group-mean difference inside the ROPE can be used to quantify the certainty that two group means are the same. One example is shown in Fig. 5.7.

To conduct the Bayesian statistical tests, we use an R statistical package named BEST [110]. To determine the ROPE on the difference between the means, we use Cohen's established

convention that the ROPE of small standardized mean difference is [-0.1,0.1] [79, 81]. Given the standard deviation being 0.3 of our data, the ROPE for the difference of means of our data is scaled to [-0.03, 0.03].

To answer question (1) about the impact of the skin type on the prediction performance, we focus on the left panel of Fig. 5.6(a). For the PD case, only 14% of the posterior distribution of the difference between the means of the lighter and darker skin groups falls in the ROPE. For the PU case, 23% of the posterior distribution falls in the ROPE. This suggests that it is highly credible to conclude that the skin type makes a difference in SpO_2 prediction, and the difference is more certain to be observed when using the back of the hand as the ROI compared to using the palm.

To answer question (2), we first focus on the left panel of Fig. 5.6(b) when participants of all skin colors are considered together. 33% of the posterior distribution of the difference of means between PU and PD cases falls in the ROPE. We then zoom into the darker skin group as shown in the left panel of Fig. 5.6(a), only 15% of the posterior distribution of the difference of means between PD and PU cases falls in the ROPE, whereas for the lighter skin group, 31% of the posterior distribution falls in the ROPE. This implies that it is highly credible that the side of the hand may have some impact on SpO_2 prediction, especially when concerning mainly the darker skin group.

5.3.3 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 5.2 shows the performance comparison of each model in leave-one-participant-out experiments. Model 1 achieved the best performance in terms of correlation and achieved the best MAE and RMSE for the PU case. Similar to the participant-specific case, the classic ratio-of-ratios method proposed in Scully *et al.* [117] achieved better MAE and RMSE results for the PD case but the correlation result was low, suggesting that the model achieved low error by simply predicting a nearly 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 model needs to accommodate for the variation in the population. As expected, in Fig. 5.6, 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.

	Hand Mode	Correlation		MAE (%)		RMSE (%)	
		Median	IQR	Median	IQR	Median	IQR
Model 1 (Proposed)	PD	**0.33**	0.42	2.33	1.07	3.07	1.52
	PU	**0.46**	0.36	**1.97**	0.80	**2.32**	0.87
Model 2 (Proposed)	PD	0.15	0.50	2.43	0.94	3.35	1.11
	PU	0.33	0.39	2.08	0.73	2.41	0.71
Model 3 (Proposed)	PD	0.23	0.38	2.48	1.18	2.98	1.33
	PU	0.27	0.31	2.02	1.03	2.54	1.28
Scully *et al.* [117]	PD	0.05	0.43	**2.08**	0.65	**2.44**	1.14
	PU	0.01	0.54	2.08	0.60	2.43	1.20
Ding *et al.* [37]	PD	0.11	0.56	3.19	1.61	3.76	1.52
	PU	0.26	0.42	2.43	1.22	2.85	1.51

Table 5.2: Performance comparison of each model structure in leave-one-participant-out experiments. Results are given as the test median and IQR of all participants.

We now revisit the two research questions raised in Section 5.3.2 under the leave-one-participant-out scenario. First, we analyze the impact of skin type given the same side of the hand. From the right panel of Fig. 5.6(a), in the PD case, only 0.04% of the posterior distribution of the difference of means between lighter and darker skin groups is within the ROPE, suggesting that the null hypothesis is rejected and the darker skin group outperforms the lighter skin group. In the PU case, 18% of the posterior distribution falls within the ROPE. This observation is consistent with the participant-specific experiments that when using the back of the hand as the ROI, the skin color is more credible to be a factor in the accuracy of SpO_2 estimation than using the palm.

Second, we analyze the impact of the side of the hand for two skin color groups. For the darker skin group shown in the right panel of Fig. 5.6(a), only 9% of the posterior distribution of the difference of means of the PU and PD cases falls in the ROPE. This shows that there is high uncertainty in the estimate of zero difference, which is consistent with the results from the participant-specific experiments. However, unlike the participant-specific experiments, for the lighter skin group, 0.2% of the posterior distribu-

tion of the difference of means between PU and PD cases falls in the ROPE. This suggests that the null hypothesis is rejected and that the PU outperforms the PD in the lighter skin group. As for the mixed group illustrated in the right panel of Fig. 5.6(b), only 8% of the posterior distribution of the difference of means falls in the ROPE, suggesting that there is a high uncertainty to conclude that PU and PD cases are comparable.

This different generalization capability in the PU and PD cases may be attributed to the 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.

5.3.4 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 the uncontrolled impact of such factors as skin tone and hair. In the first ablation study, we compare nonlinear to linear channel combinations. 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. In the second study, we

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

Table 5.3: Numerical results of the ablation studies for Model 1 (M1) in the leave-one-participant-out mode. Comparisons among the proposed (nonlinear) M1, modified M1 with only linear channel combinations, and modified M1 with fully connected dense layers instead of convolutional layers are listed. Ablation studies confirm that the nonlinear channel combinations and convolutional layers improve model performance.

compare the performance of using convolutional layers for temporal feature extraction to using fully-connected dense layers. We create this second variant of Model 1 and repeat leave-one-participant-out experiments.

Table 5.3 presents the medians and IQRs specified for numerical comparison of the ablation study. First, we compare the first and the third rows in Table 5.3 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 5.3. 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.

5.4 Discussions

5.4.1 Contact-based Dataset Testing

We also test our models on the publicly available dataset gathered by Nemcova *et al.* for their SpO_2 estimation work [98]. This dataset consists of contact-based smartphone video recordings where a participant placed a finger on the smartphone camera and was illuminated by the camera flashlight. Participants were asked to breathe normally without following any sophisticated breathing protocol. Each recording lasts about 10 to 20 seconds. The subject for each recording is not identified, so subject-specific and leave-one-participant-out experiments cannot be conducted. There is a single reference SpO_2 value associated with each recording. We used 14 recordings for training and seven recordings for testing and compared them with the modified ratio-of-ratios method proposed in their paper.

As shown in Table 5.4, Models 1 and 2 outperform the method used by Nemcova *et al.* on both the training and test recordings. Model 3 is not able to generalize well from the training set to the test set, which may be due to the small size of the dataset. It should be noted that because the participants were not asked to follow any sophisticated breathing protocol, the dynamic range of SpO_2 values is narrow. These results show that our CNN Models 1 and 2 work well for contact-based video recordings in addition to contactless video recordings.

	MAE (%)		RMSE (%)	
	Training	Test	Training	Test
Model 1	0.86	**1.19**	0.94	**1.36**
Model 2	0.50	1.28	0.59	1.64
Model 3	0.75	3.28	0.99	3.69
Nemcova *et al.* [98]	2.05	2.18	2.24	2.36

Table 5.4: Experimental results of proposed methods on the contact-based video SpO_2 dataset from Nemcova *et al.* [98]. One SpO_2 estimate was output per recording and MAE and RMSE were calculated across all recordings. Models 1 and 2 outperform the method proposed by Nemcova *et al.*, Model 3 was unable to generalize well to the test set.

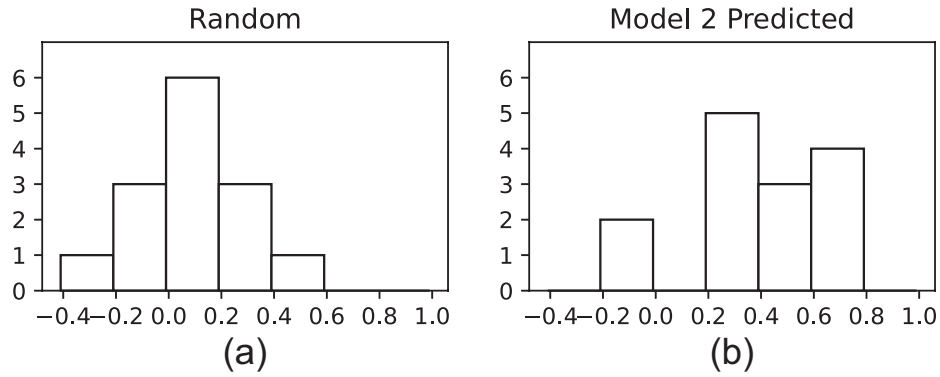


Figure 5.8: Histograms of correlation values between reference SpO_2 signals and (a) randomly generated SpO_2 signals, or (b) SpO_2 signals predicted by neural network Model 2. The correlation distribution for Model 2 is centered much higher than the random guess, confirming Model 2's capability to track SpO_2.

5.4.2 Ability to Track SpO_2 Change

By employing the standard machine learning methodology of training-validation-test split in Section 5.3 to learn neural networks that perform well on unseen data, we have already ensured the generalizability of our models [122, Chapter 11]. As further evidence that our models are capable of outputting meaningful predictions, we compare SpO_2 predictions from our learned models to randomly generated SpO_2 values. For each reference signal, a random prediction signal was generated by choosing SpO_2 values between the minimum and maximum values from the reference signal and applying a moving aver-

age window in the same way as is applied to the neural network predictions. Fig. 5.8(a) shows a histogram of the correlations between the reference SpO_2 signals and the randomly generated predictions and Fig. 5.8(b) shows a histogram of correlations between the reference SpO_2 signals and the predictions generated by Model 2. It is revealed that the neural network with a median correlation of 0.41[1] outperforms random guessing with a median correlation of -0.02, confirming Model 2's capability to track SpO_2.

5.4.3 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

[1] It has been shown in other applications that even low correlation coefficients can be meaningful. For example, in photo response non-uniformity (PRNU) work, the device used to take a photo can be predicted with correlation values below 0.1 [11].

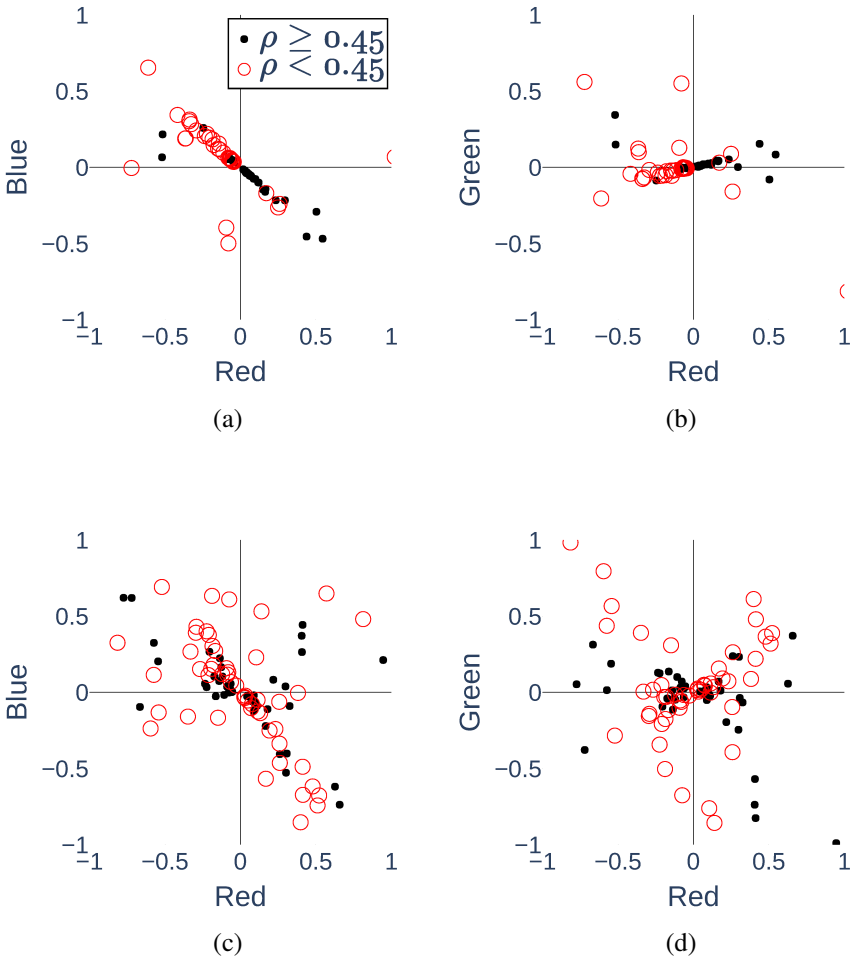


Figure 5.9: Learned RGB channel weights. Plots (a) and (b) are the channel weights learned by different model instances trained on the data of all study participants together, projected onto the RB and RG planes in the RGB space. Plots (c) and (d) are the RB and RG projections of the learned channel 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.

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. 5.9(a) and 5.9(b) 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 lines 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, and the ratio of the green channel weight to the 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 under the following setup: 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. 5.9(d) reveals that most of the better-performing instances (with $\rho \geq 0.45$) have little contribution from the green channel. In Fig. 5.9(c), 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 biophysical understanding of how light is absorbed by hemoglobin in the blood. Recall that Fig. 1.5 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.

5.5 Chapter Summary

We have proposed the first CNN-based work to solve the challenging problem of video-based remote 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 palm-up case for the lighter-skin group. 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.