Measuring Oxygen Saturation With Smartphone Cameras Using Convolutional Neural Networks

Xinyi Ding, Damoun Nassehi, and Eric C. Larson

***Abstract*—Arterial oxygen saturation (SaO_2) is an indicator of how much oxygen is carried by hemoglobin in the blood. Having enough oxygen is vital for the functioning of cells in the human body. Measurement of SaO_2 is typically estimated with a pulse oximeter, but recent works have investigated how smartphone cameras can be used to infer SaO_2. In this paper, we propose methods for the measurement of SaO_2 with a smartphone using convolutional neural networks and preprocessing steps to better guard against motion artifacts. To evaluate this methodology, we conducted a breath-holding study involving 39 participants. We compare the results using two different mobile phones. We compare our model with the ratio-of-ratios model that is widely used in pulse oximeter applications, showing that our system has significantly lower mean absolute error (2.02%) than a medical pulse oximeter.**

***Index Terms*—Convolutional neural networks, mobile sensing, oxygen saturation.**

I. Introduction

With various embedded sensors, mobile devices like smartphones and tablets have been increasingly used as out-of-the-clinic health care platforms. Such mobile health care platforms enable physiological parameters like blood pressure, heart rate, and arterial oxygen saturation (S_aO_2) to be measured more frequently, without travel to a health clinic. Re-purposing sensors on a mobile phone for health care sensing has been investigated by a number of researchers: Mehta *et al.* used the accelerometer of smartphone to detect voice disorders [1]. Kaiser *et al.* utilized the smart phone built in audio sensor to capture signals from a vortex whistle to track pulmonary function [2]. Chandrasekaran *et al.* used smartphones to estimate blood pressure [3] and Scully *et al.* used smartphones to measure oxygen saturation [4]. In this work, we also measure oxygen saturation using smartphone cameras, investigating methods that use convolutional neural networks for automatically extracting features from spatially averaged video streams of a participant's finger.

Manuscript received August 13, 2018; revised October 31, 2018 and November 16, 2018; accepted December 13, 2018. Date of publication December 17, 2018; date of current version November 6, 2019. This work was supported by the DigiDoc Technologies (Award number: 171571). *(Corresponding author: Xinyi Ding.)*

X. Ding and E. C. Larson are with the Department of Computer Science and Engineering, Southern Methodist University, Dallas, TX 75205 USA (e-mail: xding@smu.edu; eclarson@smu.edu).

D. Nassehi is with digiDoc Technologies, 4372 Egersund, Norway (e-mail: damoun@digidoctech.no).

Digital Object Identifier 10.1109/JBHI.2018.2887209

Convolutional neural networks have been used to solve many practical problems in the biomedical field. For example, measurement of various physiological signals like pupil dilation [5] and blood pressure [3]. We propose using 1D convolutional neural networks for regressing oxygen saturation. To train and evaluate our model, we conducted an IRB-approved human subjects study with 39 participants. We show that convolutional networks are able to achieve better estimates of S_pO_2 than the most common methodology, the ratio-of-ratios model [4], when measuring on mobile devices. The contributions of our work are five-fold:

1) We combine large motion detection with Singular Value Decomposition (SVD) technique to remove motion artifacts and impute sensors stream quantities, which is more robust than the previous work [6].
2) We investigate a number of different convolutional neural network architectures for their ability to measure oxygen saturation from mobile phone videos.
3) We conduct a 39 participant human study and show that our model is able to achieve significantly better S_pO_2 estimates than the ratio-of-ratios model [4].
4) We evaluate our proposed model on two different devices and show that results for each device are similar.
5) We investigate the use of different video frame rates, showing that the methods work similarly in the range from 30FPS to 240FPS.

II. Pulse Oximetry Primer and Related Work

Mobile technology touches every aspect of our daily lives, from gaming to social life. The accessibility of mobile devices makes them a natural choice for health care delivery. Based on a report from Centers for Medicare and Medicaid Services (CMS) [7], the U.S. health care costs were $3.3 trillion in 2016, which equals 17.9 percent of GDP that year and it is projected that spending will grow 5.3 percent per year until 2024. The U.S. health care system could save up to $7 billion a year by using mobile health apps according to IQVIA [8]. Mobile health apps could provide users more immediate and more frequent feedback about their health conditions and, thus, have gained popularity in recent years. Pulse oximetry measurement from mobile phones is no exception to this trend.

Oxygen is vital for the functioning of cells in our body. Normally, the oxygen saturation of a healthy individual is above 95%. If the oxygen saturation level drops below 95%, it is a strong indicator of oxygen delivery imbalance [9]. Extremely

Fig. 1. Experimental setup of ground truth pulse oximetry data collection and custom phone application.

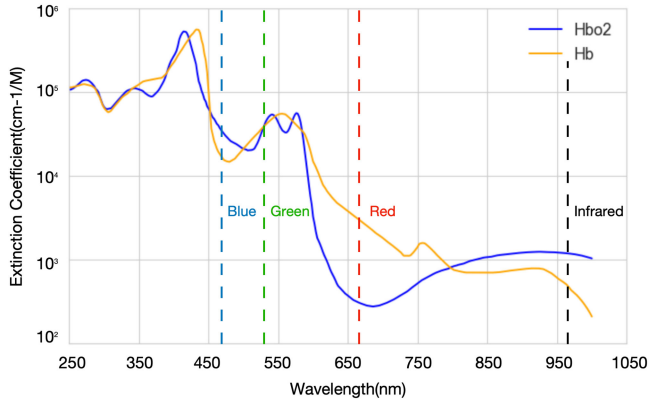

Fig. 2. Extinction coefficients for different wavelength lights.

low oxygen saturation could lead to hypoxia. However, oxygenation can be used as a indicator for other diseases. For instance, oxygen delivery imbalance could be the result of diseases like pneumonia and asthma, for which pulse oximetry can be used to differentiate them from less severe illnesses like the common cold [9].

Hemoglobin is a protein that helps deliver oxygen molecules to our body through the circular system. One hemoglobin cell can carry up to four oxygen molecules and, when this occurs, it is called oxyhemoglobin (HbO_2). When there are fewer than four molecules, it is called deoxyhemoglobin or reduced hemoglobin (Hb). The arterial oxygen saturation S_aO_2 in blood can be estimated using the following equation:

$$S_aO_2 = \frac{c_{HbO_2}}{c_{HbO_2} + c_{Hb}} 100\%$$

Here, c_{HbO_2} is the concentration of oxyhemoglobin and c_{Hb} is the concentration of deoxyhemoglobin. Although there are other kinds of hemoglobins like methemoglobin and carboxyhemoglobin, their concentration levels are sufficiently low and can be safely ignored when calculating the S_aO_2.

The gold standard for measuring S_aO_2 is a gas chromatograph [10]. However, this process is invasive and requires a blood sample, which makes it unsuitable for continuous measurement. Takuo Aoyagi [11] first proposed exploiting the pulsation of arterial blood to measure oxygen saturation, thus this technique is called pulse oximetry and the measurement device is called a pulse oximeter.

Pulse oximeters are the most widely used devices for measuring oxygen saturation in hospitals, critical care units, and homes. The measurement from a pulse oximeter is typically written as S_pO_2. S_pO_2 and S_aO_2 are highly correlated, usually with discrepancy less than 3% provided the S_aO_2 is above 70% [11]. The most widely implemented model used in pulse oximeters is called the ratio-of-ratios model [11]. In this model, two LEDs are shown onto the finger at 660 nm red light and 940 nm infrared light with detectors on the opposite side of the finger. Oxyhemogoblin and deoxyhemoglobin have different extinction coefficient rates for these two wavelengths as shown in Fig. 2. By analyzing the light incidents measured by each detector, oxygen saturation level can be estimated. We briefly derive the ratio-of-ratios model here using the same notation as used in [11].

During systole, the arteries have increased blood compared to diastole. When shining light on the finger, this results in changing the path that light travels through before encountering the detector. The light path consists of two biological areas [11]: The first area we denote as d_{DC} which does not change in terms of time t. This area mainly contains tissue, venous blood, and non-pulsatile arterial blood. The second area is the pulsatile arterial blood which changes in terms of time t. Let $d_{\min}$ denote the minimum path of the pulsatile arterial blood the light has to pass through (diastole) and $d_{\max}$ denote the maximum path (systole). Using Beer-Lambert's law, we have

$$I_H = \alpha e^{-\beta d_{\min}}$$

$$I_L = \alpha e^{-\beta d_{\max}}$$

where

$$\alpha = I_0 e^{-\varepsilon_{DC}(\lambda) c_{DC} d_{DC}}$$

$$\beta = \varepsilon_{Hb}(\lambda) c_{Hb} + \varepsilon_{HbO_2}(\lambda) c_{HbO_2}$$

where I_0 is the incident light and λ refers to the wavelength of light. During diastole, the path is minimal, thus the transmitted light has large amplitude (I_H). During systole, the transmitted light has small amplitude (I_L). If we let $\Delta d = d_{\max} - d_{\min}$, then the light intensity that reaches the detector can be expressed as:

$$I = I_H e^{-\beta \Delta d}$$

The ratio of absorption of two lights is given by R in equation (1). If we assume the two different lights have the same path length through the finger then the Δd cancels and we can rearrange the equation to get equation (2), containing the S_aO_2 term. This assumption simplifies the derivation of the ratio-of-ratios model. In practice, a calibration process is usually required.

$$R = \frac{A_{t,660}}{A_{t,940}} = \frac{\ln(I_{L,660}/I_{H,660})}{\ln(I_{L,940}/I_{H,940})} \quad (1)$$

$$R = \frac{\varepsilon_{Hb}(\lambda_{660}) + [\varepsilon_{HbO_2}(\lambda_{660}) - \varepsilon_{Hb}(\lambda_{660})] S_a O_2}{\varepsilon_{Hb}(\lambda_{940}) + [\varepsilon_{HbO_2}(\lambda_{940}) - \varepsilon_{Hb}(\lambda_{940})] S_a O_2} \quad (2)$$

Petersen *et al.* [9] applied this ratio-of-ratios model using an external sensor that connected to a mobile phone headphone jack. The Phone-based oximeter required a conventional clinical oximeter finger sensor, which sent collected data through the audio headset interface to the smartphone. The smartphone was used as a data analysis platform.

Scully *et al.* [4] used a similar model, given in equation (3) to estimate the oxygen saturation using a mobile phone camera. Their system did not require the clinical sensor and, instead, used the built-in flash and red/blue channels of the camera. The main difference is that they used the blue channel instead of infrared and normalize each measure by the DC value.

$$S_pO_2 = A - B\frac{AC_{RED}/DC_{RED}}{AC_{BLUE}/DC_{BLUE}} \tag{3}$$

where A and B must be found through calibration. The ratio-of-ratios model simplifies the situation by ignoring the light scattering issue, thus a calibration process is required [11]. Reddy *et al.* [12] proposed a calibration free model. Their model assumes portions of the signal to be linear, which is an optimistic assumption and usually results in reduced performance in practice.

One major issue when estimating oxygen saturation from a finger is motion artifacts that mask the signal differences between systole and diastole. Yadhuraj *et al.* [13] compared different motion artifact removal techniques and found two techniques that usually work better in practice: Singular value decomposition (SVD) [6] and Fourier series analysis [14]. Both techniques attempt to recreate the signal cycle-by-cycle. In our work, we modify the SVD method to make it more robust to large motion artifacts, detailed later on.

Researchers have also investigated remote oxygen saturation measurement techniques [15]–[17]. Their methods typically employ detecting a user's face and using the RGB camera and ambient light to estimate differences between systole and diastole on the surface of the skin. However, these remote measurement techniques require the user to sit in a fixed position and have distance limitations. Moreover, the signals are often more noisy, resulting in decreased performance.

III. Experimental Procedures

To evaluate our model, we conducted an human subjects study involving 39 participants (male = 27, female = 12, ag e = 18–30), approved by university IRB, Study ID H17-146-LARE. All participants were college students who identified as not having pulmonary or heart diseases. Fig. 1 shows the set up of the experiment with a custom iPhone data collection application and ground truth pulse oximeter. All experiments were conducted in a controlled room on campus.

During the experiment, participants sat in a comfortable chair with all equipment on a desk in front of them. Optionally, the participant could elect to perform some physical activity before the experiment (such as jumping jacks), which can help to deoxygenate the blood more quickly when the participant holds their breath. The participant took the following measurements:

Baseline measurement: A pulse oximeter (Nellcor PM10N [18]) was clamped on the participant's right index finger. The pulse oximeter reports the S_pO_2 once per second and has an accuracy of $\pm 2\%$ when the oxygenation level is above 70%. At the same time, the camera of an iPhone 6 s was placed on the right middle finger and the white LED torch was turned on. A black cloth was placed in between the middle and index finger to prevent light transference between the oximeter and torch. The brightness of the white LED was set to 20% to prevent overheating (found by trial and error for keeping the torch lit for extended periods without discomfort). The participant was asked to keep their hand as still as possible for the duration of the experiment. The participant then breathed normally for 30 seconds. Oxygen saturation and heart rate were collected during this time.

Breath-Holding: After 30 seconds, the participant was asked to hold their breath as long as they could, to the point that it caused some discomfort. The participant was asked not to move while holding their breath as this can influence the accuracy of the oxygen measurement. Once the participant determines that they cannot hold their breath any longer, they are asked to breath normally until they feel ready to repeat the procedure. The participant was allowed to take a break at anytime (or discontinue if they felt overly fatigued).

The above baseline and breath-holding experiments were conducted three times for each participant. After completing three iterations for the iPhone 6 s, we repeat the above procedure using an iPhone 7 Plus. The reason for doing this is because we wanted to investigate the effect of the longer distance between the camera and flash of the iPhone 7 Plus. The iPhone 7 plus has two cameras, telephoto and wide-angle. We only use the telephoto because it is closer to the white LED. For the iPhone 7 Plus, we changed the brightness level of the torch to 60% to increase strength signal. The iPhone 7 Plus can sustain 60% torch brightness without discomfort (compared to the iPhone 6 s that required 20% brightness; both measure found through informal trial and error). We set the frames per second (FPS) of the camera to the maximum allowed by the firmware, which is 240 for both iPhone 6 s and iPhone 7 Plus. The recording format was also set to use lossless compression. However, due to limited computing resources and the time required to stream the lossless video to file, the FPS could not maintain 240FPS while running our application. The actual frame rate recorded was about 220FPS for the iPhone 6 S, on average, and about 235FPS for the iPhone 7 Plus. To eliminate issues with frame rate, we interpolate to 240 FPS using linear interpolation before any further processing.

This session, which includes consenting the participant and data collection, took approximately 30 minutes to complete. The participant was asked to complete the entire session twice on two different days. For involvement in the study, we gave participants a $10 gift card for finishing both sessions. For iPhone 6 s, we recorded 39 participants' data, resulting in a total of 34,716 seconds. The average of each iteration was 156 seconds. For the iPhone 7 Plus, we recorded 37 participants' data, with total 32,521 seconds. The average of each iteration was 154 seconds. Fig. 3 shows the distribution of all S_pO_2 values recorded during

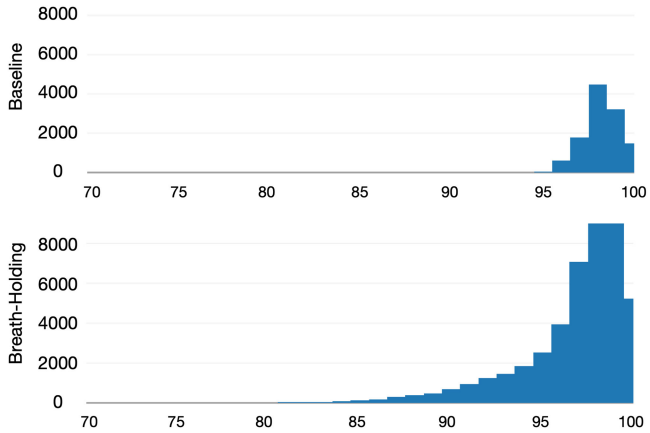

Fig. 3. SpO_2 distribution of all participants.

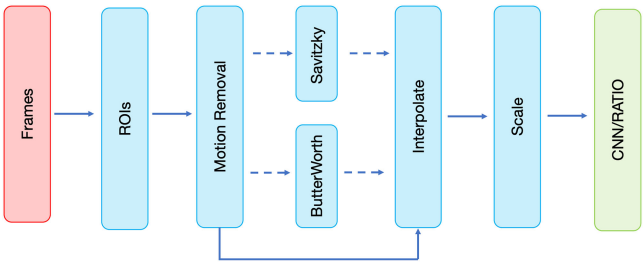

Fig. 4. Flowchart of data processing.

the experiments from the ground truth pulse oximeter (reported at once per second). Counts are grouped by whether they are taken during baseline measurement or during breath-holding. Most values during the baseline period are above 95%. During breath holding, the values varies from 73% to 100%. However, we noticed the recovery is very fast once the participant stops holding their breath, so there are not so many low S_pO_2 values in the histogram.

IV. Method

Fig. 4 gives an overview of our data processing pipeline. We first convert RGB frames to photoplethysmography (PPG) signals by averaging pixel values from candidate regions of interest (ROIs). Each ROI is a square area in the frame of the video. Two competing approaches are investigated for calculating the PPG signal from candidate ROIs. The first approach is to average pixel values from the middle 100 × 100 window. The second approach is to first divide the frame into 16 × 16 different ROIs (320 × 180 per ROI). The final signal is a weighted average of the different ROIs, where each weight is calculated from the signal to noise ratio (SNR) of a given ROI, as follows:

$$PPG_{t,i} = \frac{SNR_{t,i}}{\Sigma_i SNR_{t,i}} ROI_{t,i} \tag{4}$$

$$SNR_{t,i} = \frac{\mu_{ROI_{t,i}}}{\sigma_{ROI_{t,i}}} \tag{5}$$

Our weighted averaged method is similar to that used by Gauzzi *et al.* [15]. However, we use a simpler SNR equation because we notice the texture of fingertip is relatively uniform compared to captured videos from the face used in [15]. We compared these two approaches showing that they perform approximately equal (discussed further in the result section).

A. Motion Artifact Removal

Even though the participants are asked to remain still during data collection, we notice there are many motion artifacts, especially during the seconds immediately before the oxygen saturation level is lowest. This is understandable because the participant is maximally discomforted during these seconds. Based on a survey paper [13], there are two popular methods for motion artifact removal from the PPG: Fourier analysis [14] and Singular Value decomposition (SVD) [6]. Both methods try to recreate the new signal cycle-by-cycle. The Fourier analysis method requires a good estimate of the heartrate period, which is impractical when there is a considerably large motion. As such, we decided to use the SVD method and modify it to make it more robust to large motion artifacts. Intuitively, the idea of SVD method is to create a matrix where each row is a period of the PPG signal. Ideally, the rank of this matrix would be singular because all subsequent rows are repetitions of the first (and thus only one eigenvalue of the matrix would be non-zero). The specifics of the SVD methods are as follows: we first reshape the signal into a matrix with row length approximately equal to the period,

$$X = \begin{bmatrix} x(1) & x(2) & \dots & x(n) \\ x(n+1) & x(n+2) & \dots & x(2n) \\ \vdots & \vdots & \dots & \vdots \\ x(nm-n+1) & x(mn-n+2) & \dots & x(mn) \end{bmatrix}$$

In the ideal situation, the first eigenvalue of the matrix will be nonzero and all other eigenvalues will be zero. However, because of noise and slight irregularities in the PPG signal, all eigenvalues will be nonzero. Thus, we vary the row length n and calculate the ratio of the first two eigenvalues. When this ratio is maximized, n should approximately match the actual period of the heart rate.

When applying the SVD method, there is a fixed size window (8 seconds for example) working like a queue. The window of data is reshaped to a matrix and the best row length, n, is found. We then average all the rows in the matrix to create a "new" cycle, and replace the first n samples with this average cycle. Finally, n samples are removed from the queue and n new samples are added to the end of the queue, and the process repeats. In this way, motion artifacts are replaced by the average of cycles in a window.

However, we notice that "large" motion artifacts result in recreated cycles that are not natural as shown in Fig. 5. The first row is the raw signal, with a clear large motion artifact around 71 seconds. The second row is the recreated signal using the SVD method. In this example, we can see that the sudden motion artifact biases the whole signal; after the motion, a new baseline measure occurs above the original bias. This causes an unnatural replacement cycle from averaging. We modify SVD to resolve this unnatural signal. In our method, we first detect

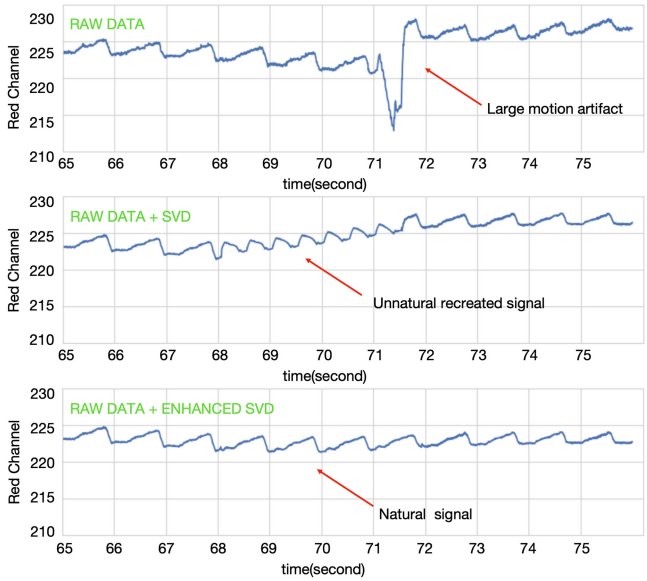

Fig. 5. Comparison of signals using different artifact removal methods. Bottom: our proposed method.

large motion artifacts and then shift the signal by a margin. To detect the motion artifacts, we use a similar method as in [19]: we first calculate the average gradient of the PPG signal in half second windows. If two consecutive windows' average gradient differs by more than 50%, we flag that window as containing a motion artifact. We then calculate the mean difference between the previous and immediately following five heart rate cycles. We shift the signals after the motion artifact up/down by this mean difference, effectively removing any sustained shift in the bias of the signal due to motion. Furthermore, we alter the process for replacing cycles from the original SVD method. When constructing the signal matrix, if we detect there are motion artifacts in a coming cycle, we do not add this amount of data to the SVD matrix but use the following same amount of data before calculating the average. We perform this simplification because we notice that neighboring cycles are typically similar in shape and it is difficult to recover signals having large motion artifacts. As we can see in Fig. 5 (bottom), using our method, the recreated signal is visibly more natural.

B. Filtering

The resulting PPG signal includes a tremendous amount of information. When processing the PPG signal for heart rate or or heart rate variability, we usually apply a bandpass filter to detrend the signal and the signal is often inverted [19], [20]. However, we observe that, for the red channel, the trend of the PPG signal decreases as the oxygen saturation level decreases. This agrees with the fact that as the oxygen saturation decreases, there will be more deoxyhemoglobin and less oxyhemoglobin. Deoxyhemoglobin absorbs more red light than oxyhemoglobin, which results in less red light that reaches the camera. This makes us hypothesize that this bias information is also important for estimating the S_pO_2. Therefore, we decompose the signal into bandpass and lowpass filtered versions. We calculate the

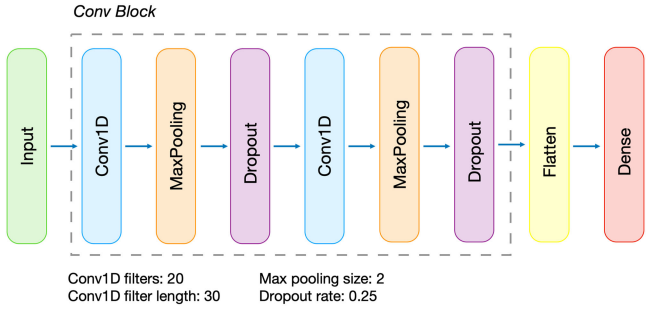

Fig. 6. 1D Convolutional Neural Network Architecture. Best parameters found through grid search.

bandpass version by applying a Butterworth bandpass filter from 0.7 HZ to 4 HZ, traditionally used in the calculation of the PPG signal. The bias is calculated through lowpass filtering the signal using a Savitzky-Golay filter (10 seconds, order 3). We run experiments with these signals used as input features, investigating if there is any advantage to using them separately, combined, or using the unfiltered raw signal. Before sending these data into the model, they are interpolated to be 240 samples per second and scaled to have zero mean and unit standard deviation. We do not invert the signal, because we believe the convolutional neural network can learn such information from the input data.

C. Convolutional Network Architecture

After preprocessing, we have a raw PPG signal, bandpass filtered signal, and bias signal for each color channel. In order to extract features automatically from these signal streams, we investigate the use of Convolutional Neural Networks or CNNs. CNNs are feedforward neural networks with filtering layers that are learned via back-propagation. Each convolutional layer applies a multiple convolution operations to the input, concatenates the results, and passes these results to the next layer. More interested readers can refer to [21]. Fig. 6 gives the overview of our 1-D convolutional neural network model. Compared with the ratio-of-ratios model that only uses two wavelengths of light, we input all three channel data, keeping as much information as possible. Our network consists of two temporal convolutional layers comprised of max pooling and 25% dropout (to help mitigate overfitting). The number of convolutional layers was investigated as a hyperparameter. The input window size, filters, and filter length are also considered hyperparameters. We use a cross validated grid search process to find the best combination of hyperparameters (a total of 3800 parameter combinations were investigated). The best hyperparameters are shown in Fig. 6. We use both dropout and early stopping to prevent overfitting and optimize our model in mini-batches using adaptive momentum (ADAM) [22]. We implemented our convolutional neural network models using Python 3.5 and Tensorflow 1.2 [23]. We trained our models on a cluster of 36 nodes each having 256 GB memory and accelerated by an NVIDIA P100 GPU. Total training time for the grid search took approximately 4 weeks.

TABLE I
OVERALL RESULTS COMPARISON

Model	iPhone 6S		iPhone 7 Plus	
	RMSE	MAE	RMSE	MAE
Ratio-of-Ratios	3.07	2.52	3.40	2.80
Conv (Raw PPG)	**2.54**	**2.02**	**3.01**	2.38
Conv (BW)	2.74	2.08	3.07	2.40
Conv (BW + Bias)	2.68	2.05	3.03	**2.33**
Conv (Raw PPG) + Augmentation	2.81	2.19	3.20	2.51

V. RESULTS

For each model, we cross validate training and testing data using leave-one-participant-out for each phone. That is, when testing on a given participant, we never use their data in training the models and we use training and testing data within a particular device (6s or 7 Plus). Table I gives the overall results comparison of different models and devices averaged over the 39 participants (37 participants for iPhone 7 Plus). When comparing the ground truth pulse oximeter to the predicted S_pO_2, we report the root mean square error (RMSE) and mean absolute error (MAE).

As mentioned, we also investigate using different input streams for the convolutional architecture: Raw PPG, Butterworth bandpass filtered (BW), and Butterworth bandpass filtered with bias term from a Savitsky-Golay filter (BW + Bias). Moreover, we augment the training data using an oversampling technique during training, which will be explained in next subsection. In the raw PPG case, the model learns from the three RGB channels only, but in the BW+Bias case, it learns from 6 channels (three channels for the BW filtered signal and three channels for the bias). We can see the results of using the Butterworth filtered PPG signal (BW) is typically the worst convolutional network performer, but the difference is not significant. All convolutional approaches perform similarly—however the Raw PPG and BW+Bias signals are more consistent which supports a conclusion that separating the Bias and bandpass components of the signal is unnecessary. Thus, we suggest using the Raw PPG signal to reduce computational complexity and let the model optimize a strategy for extracting filter-based features automatically.

Also apparent from Table I, the estimates from the iPhone 7 Plus are slightly worse than from the iPhone 6 s. This is true for both the ratio-of-ratios model and convolutional models. Even though we use the telephoto camera, which is closer to the white LED, the signal is still not as strong as that from the iPhone 6 s. However, the convolutional model is better than the ratio-of-ratios model (based on an F-test of residual variance, $p < 0.01$) and the mean absolute error is within an acceptable range compared with the pulse oximeter that has $\pm 2\%$ error from the S_aO_2.

Fig. 7 shows a box and swarm plot comparison of RMSE and MAE for all users from the ratio model and convolutional model (Raw PPG). We observe there are more outliers on the iPhone 7 Plus. The measures for the convolutional architecture are visibly better than the ratio-of-ratios model and the difference is significant based on an F-test of the residual variance ($p < 0.01$).

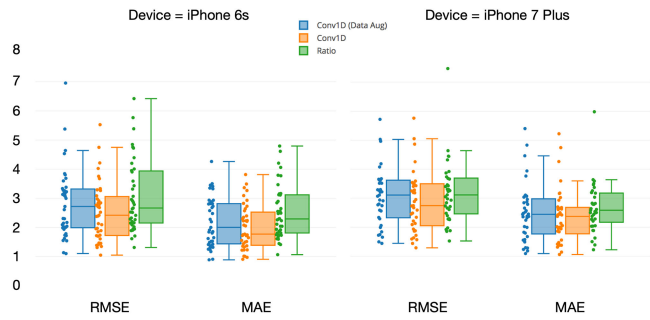

Fig. 7. Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) comparison of iPhone 6 s and iPhone 7 Plus using the convolutional model (Raw PPG) and the ratio-of-ratios method.

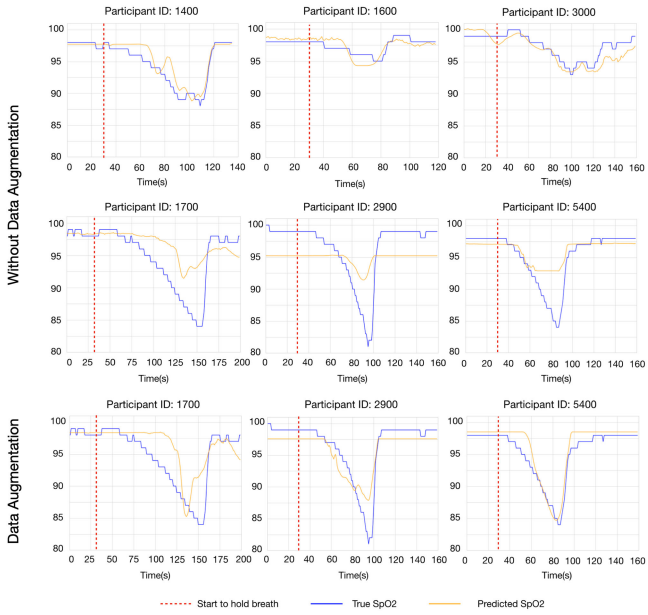

Fig. 8. Convolutional (Raw PPG) model prediction results from different participants. Top: predictions from three participants when oxygen level is above 85%. Middle: predictions from three participants when oxygen saturation level goes below 85% Bottom: predictions from the same three participants when oxygen level goes below 85% but with data augmentation.

Fig. 8 top and middle row gives six iterations of breath-holding from six different participants. We can clearly see that when the participants start to hold their breath at around 30 seconds, the oxygen saturation level does not decrease immediately. It takes different amount of time to drop for different participants, which is due to how quickly oxygen is consumed by each participant. In the top row, we can see, our model follows the true value quite well when the oxygen saturation level is above 85%. However, in the middle row, when the oxygen saturation level drops below 85%, our CNN model tends to predict higher S_pO_2 values. One reason for this is because of our breathe-holding experimental setup, it is difficult to obtain training data below 85%. This could be solved by using a more controlled hypoxic environment but greatly increases the risks associated with the experiment.

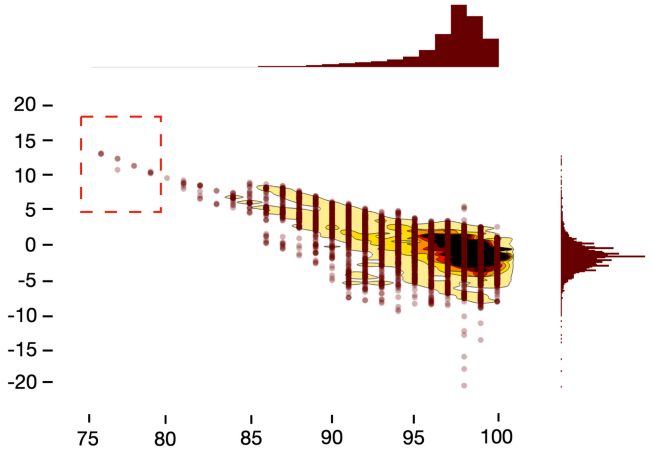

Fig. 9. Modified Bland-Altman plot for all participants using the iPhone 6 S after data augmentation. Red dash boxed points are from one single participant.

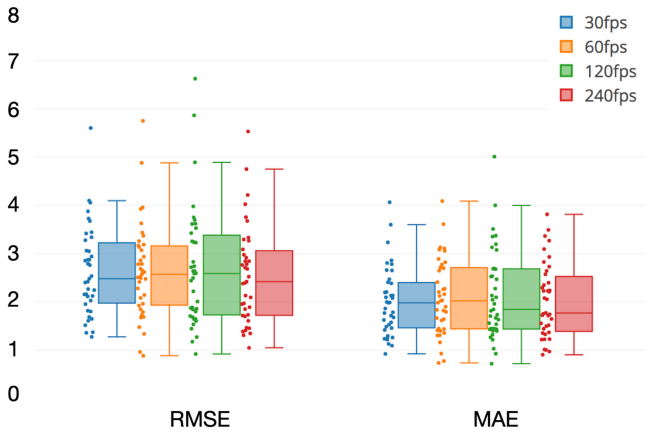

Fig. 10. Convolutional (Raw PPG) model prediction using different FPS.

One interesting observation is from Participant ID 3000 shown in Fig. 8 top right. At around 110 seconds, there is an increase in the oxygen saturation level, which means the participant may have taken a breath involuntarily and the convolutional model is able to capture this.

A. Data Augmentation

In previous analyses the predictions of our current model bias towards normal values because there are fewer "low oxygen" training points. Due to the biased nature of the data we are collecting, it is difficult to create a truly unbiased training dataset. Thus, we created a nearly unbiased dataset by 1) Reducing the baseline data to 10 seconds instead of using the entire 30 seconds. 2) We calculate the portions of data in each period and augment by reusing part of iterations while maintaining the time series properties. For example, if there are fewer training points in period 85–90, then, we augment our data by reusing (oversampling) a continuous part of an iteration that it is approximately in this range. Moreover, we used data from different participants for augmentation, avoiding some participants from being vastly overrepresented in the training data.

We only augment our data for training folds—the testing data is not manipulated. As we can see from Fig. 8, bottom, our model is able to predict oxygen saturation levels below 85% more consistently. However, we also notice the RMSE and MAE increased to 2.81% and 2.19% as listed in Table I. As discussed above, the data points below 90% include more motion artifacts because participants are more likely to move while breathe-holding, especially when the experiment become uncomfortable. Thus by oversampling these training points, we are also including more motion artifacts. This performance decrease is more severe for iPhone 7plus when the signal is already noisy before augmentation.

Fig. 9 shows the modified Bland-Altman Plot of all participants of our best convolutional model for iPhone 6 s (the x-axis is not the mean of the two devices, but is instead the ground truth measurement) with data augmentation. Because the training labels from pulse oximeter are integer values only, we can see there are some gap between points in the plot. The 95% prediction intervals (PI) for the plot are as follows ($\mu \pm 1.96 \times \sigma$):

- For all points the PI is $(-7.02, 4.04)$.
- When S_pO_2 is 95%—100% the PI is $(-6.49, 2.43)$
- When S_pO_2 is 90%—95% the PI is $(-6.01, 5.85)$.
- When S_pO_2 is 85%—90% the PI is $(-2.22, 9.22)$.

As we can see, the augmented model tends to predict increased values compared to the ground truth when the oxygen saturation level drops below 90%. However, without data augmentation these prediction intervals are noticeably worse for low oxygen saturation levels. The red dash boxed points are from one single participant. It is the only participant that has oxygen saturation level below 80% in our dataset. Since we are using leave one subject out cross validation, there are no training points below 80% when predicting this user, which explains the large prediction errors.

B. Compare Different FPS

Modern smart phones allow video recording at high frequency to support "slow motion" capture. When designing our experiment, we exploit this advantage, hoping to get as much data as possible for our CNN model. However, it is unclear if this high FPS recording is advantageous for the CNN model—it is possible that there might be simply redundant data when capturing at a high FPS. As a further analysis, we downsampled our 240FPS signal to 30FPS, 60FPS and 120FPS and trained new CNN models for each framerate. Boxplots of the RMSE and MAE are given in Fig. 10. We observe similar boxplots regardless of the FPS. This supports a conclusion that the CNN model can achieve similar predictions even if the capture rate is only 30FPS.

C. Discussion

All the results reported are from the estimate of the PPG calculated from the center 100×100 ROI. When comparing this estimate to the weighted average of ROIs, the result are nearly identical. We hypothesize this is because the texture distribution

of the fingertip is quite uniform, such that each ROI is likely to include the same amount of information. Thus, weighted averaging of ROIs is unnecessary when capturing data from the finger.

Signal processing techniques like motion artifact detection and filtering are computationally intensive, especially for mobile devices. Modern smart phones are equipped with a number of sensors, including an accelerometer and gyroscope. The built in accelerometer on modern smartphones is fairly sensitive and could capture motion artifacts independent of the camera. Such processing methods could be key for performing motion artifact removal in real-time.

One limitation of convolutional networks is that they require many example data to properly train. Because we train a CNN for each smartphone type, it is unclear if a generalizing CNN model can be trained that generalizes across smartphones. We leave this analysis to future work.

VI. Conclusion

In this paper, we conducted a systematic analysis of measuring oxygen saturation on mobile phones with convolutional neural networks. We combined SVD motion artifact removal with motion detection to create more natural signals at extreme conditions. We conducted a breath-holding human subjects experiment involving 39 participants to evaluate our model. We compared the results from two different mobile phones that have different distances between the camera and light source. Using leave-one-participant-out cross validation, our model is able to achieve better results than the commonly used ratio-of-ratios method, with mean absolute error of 2.02% compared to a medical pulse oximeter.

References

[1] D. D. Mehta, M. Zanartu, S. W. Feng, H. A. Cheyne II, and R. E. Hillman, "Mobile voice health monitoring using a wearable accelerometer sensor and a smartphone platform," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 11, pp. 3090–3096, Nov. 2012.

[2] S. Kaiser *et al.*, "Design and learnability of vortex whistles for managing chronic lung function via smartphones," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, 2016, pp. 569–580.

[3] V. Chandrasekaran, R. Dantu, S. Jonnada, S. Thiyagaraja, and K. P. Subbu, "Cuffless differential blood pressure estimation using smart phones," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 4, pp. 1080–1089, Apr. 2013.

[4] C. G. Scully *et al.*, "Physiological parameter monitoring from optical recordings with a mobile phone," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 2, pp. 303–306, Feb. 2012.

[5] C. Wangwiwattana, X. Ding, and E. C. Larson, "Pupilnet, measuring task evoked pupillary response using commodity RGB tablet cameras: Comparison to mobile, infrared gaze trackers for inferring cognitive load," *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.*, vol. 1, no. 4, 2018, Art. no. 171.

[6] K. A. Reddy and V. J. Kumar, "Motion artifact reduction in photoplethysmographic signals using singular value decomposition," in *Proc. IEEE Instrum. Meas. Technol. Conf.*, 2007, pp. 1–4.

[7] "Historical" Centers for Medicare & Medicaid Services, Baltimore, MD, USA. Accessed: May 17, 2018. [Online]. Available: https://tinyurl.com/cm5jfk4

[8] "Medicine use and spending in the U.S. A review of 2017 and outlook to 2022," IQVIA, Durham, NC, USA. Accessed: May 17, 2018. [Online]. Available: https://tinyurl.com/ydatt7bx

[9] C. L. Petersen, T. P. Chen, J. M. Ansermino, and G. A. Dumont, "Design and evaluation of a low-cost smartphone pulse oximeter," *Sensors*, vol. 13, no. 12, pp. 16882–16893, 2013.

[10] Radiometer, "ABL800 FLEX reference manual," Radiometer Medical ApS, 2012.

[11] J. G. Webster, *Design of Pulse Oximeters*. Boca Raton, FL: CRC Press, 1997.

[12] K. A. Reddy, B. George, N. M. Mohan, and V. J. Kumar, "A novel calibration-free method of measurement of oxygen saturation in arterial blood," *IEEE Trans. Instrum. Meas.*, vol. 58, no. 5, pp. 1699–1705, May 2009.

[13] S. Yadhuraj and H. Harsha, "Motion artifact reduction in photoplethysmographic signals: A review," *Int. J. Innovative Res. Develop.*, vol. 2, no. 3, pp. 626–640, 2013.

[14] K. A. Reddy, B. George, and V. J. Kumar, "Use of fourier series analysis for motion artifact reduction and data compression of photoplethysmographic signals," *IEEE Trans. Instrum. Meas.*, vol. 58, no. 5, pp. 1706–1711, May 2009.

[15] A. R. Guazzi *et al.*, "Non-contact measurement of oxygen saturation with an RGB camera," *Biomed. Opt. Express*, vol. 6, no. 9, pp. 3320–3338, 2015.

[16] L. Kong *et al.*, "Non-contact detection of oxygen saturation based on visible light imaging device using ambient light," *Opt. Express*, vol. 21, no. 15, pp. 17464–17471, 2013.

[17] D. Shao *et al.*, "Noncontact monitoring of blood oxygen saturation using camera and dual-wavelength imaging system," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 6, pp. 1091–1098, 2016.

[18] "Nellcor PM10N portable SPO_2 patient monitoring system infosheet," Medtronic, Minneapolis, MN, USA. Accessed: Jun. 17, 2018. Available: https://tinyurl.com/y7tq9xrh

[19] R.-C. Peng, X.-L. Zhou, W.-H. Lin, and Y.-T. Zhang, "Extraction of heart rate variability from smartphone photoplethysmograms," *Comput. Math. Methods Med.*, vol. 2015, 2015, Art. no. 516826.

[20] S. Lopez and R. Americas, "Pulse oximeter fundamentals and design," Freescale Semiconductor Inc. Application Note Document Number AN4327 Rev. 2, 2011.

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[22] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, arXiv:1412.6980.

[23] M. Abadi *et al.*, "Tensorflow: A system for large-scale machine learning," in *Proc. 12th USENIX Conf. Operating Syst. Des. Implementation*, 2016, pp. 265–283.