WristO2: Reliable Peripheral Oxygen Saturation Readings from Wrist-Worn Pulse Oximeters

Caleb Phillips, Daniyal Liaqat, Moshe Gabel and Eyal de Lara
Department of Computer Science
University of Toronto
Toronto, Canada
{caleb, dliaqat, mgabel, delara}@cs.toronto.edu

Abstract—Peripheral blood oxygen saturation (SpO2) is a vital health signal with many clinical applications. Modern wrist-worn devices, such as the Apple Watch, FitBit, and Samsung Gear, have pulse oximeter sensors, making them theoretically capable of measuring SpO2. However, current techniques for SpO2 measurements using pulse oximeter sensors are based on readings taken from the fingertip. Readings collected from the wrist are unreliable and often inaccurate, due to motion and insufficient skin contact. Enabling accurate oxygen saturation monitoring on wearable devices would allow continuous health monitoring and open up new avenues of research.

In this work, we explore the reliability of SpO2 measurements from the wrist. Using a custom wrist-worn pulse oximeter, we find that existing algorithms used in traditional fingertip SpO2 sensors are a poor match for taking measurements from the wrist and can lead to over 90% of readings being inaccurate. We further show that skin tone, IMU sensors, and user-level calibration affect measurement error, and must be considered when designing wrist-worn SpO2 sensors and measurement algorithms.

Next, based on our findings, we propose WristO2, an alternative approach for reliable SpO2 sensing. By selectively pruning unreliable data, WristO2 achieves an order of magnitude reduction in error compared to existing algorithms, while still providing sufficiently frequent readings for continuous health monitoring.

Index Terms—Health care, Pervasive computing, Sensors

I. INTRODUCTION

Peripheral oxygen saturation (SpO2) is a measure of the percent of oxygenated blood, and its usefulness extends across domains such as sleep apnea diagnosis [1], monitoring oxygen therapy results for COPD patients [2], and patient recovery monitoring in the ICU [3]. It is also a critical measure for monitoring patients with COVID-19 [4]. Enabling frequent, unobtrusive, and reliable ambulatory monitoring of oxygen saturation could be a game changer in such domains, for example by allowing early interventions that could drastically improve health outcomes and reduce health care costs [5].

However, current approaches for home SpO2 monitoring only provide intermittent readings since they require active user interaction. To measure their SpO2 levels, users must press their fingertip closely to the sensor for 30 seconds at a time. As part of the growing mobile health monitoring movement, some smartphone manufacturers have provided an built-in pulse oximeter on the back of smartphones (e.g. Samsung Galaxy S81) that requires the user to press a fingertip against the sensor to obtain an SpO2 reading.

Wrist-worn smartwatches such as the Apple Watch, FitBit, and Samsung Gear already contain pulse oximeters albeit they tend only to be used to derive heart rate. Interestingly, these pulse oximeters in theory could also be used to extract SpO2. Additionally, the device is in constant contact with the user’s skin which eliminates the need for active interaction. In practice, however, SpO2 readings from the wrist are notoriously inaccurate. While the pulse oximeters sensors on these devices are fundamentally the same as the ones used in hospital and commercial fingertip SpO2 monitors, calculating oxygen saturation from a wrist-worn sensor leads to unreliable measurements due to poorly-fitting devices, wrist and arm motion, low blood perfusion, interference from ambient light, motion, and the effects of skin tone [6]–[9]. For example, while the recently released Apple Watch 6 uses a pulse oximeter to provide SpO2 readings, users have found the measurements to be unreliable2 and its fine print asserts it is not intended for clinical purposes.

Despite the fact that most pulse oximeter readings from a wrist-worn device are unreliable, we hypothesize that occasionally, readings taken from such a device will be sufficiently reliable. Consider a patient that currently tracks her oxygen saturation twice per day using an at-home fingertip sensor kit. If she can use her smartwatch to identify even a single reliable SpO2 reading every ten minutes, we have removed the need for active user interaction and succeeded in increasing the amount of available data by almost two orders of magnitude. Therefore, even if only a small fraction of oxygen saturation readings are reliable, as long as they can be confidently identified among a majority of noisy readings we can improve the overall usefulness of wrist-worn oxygen saturation monitors.

Our Contributions: In this work, we demonstrate that a reliable SpO2 signal can be taken automatically from the wrist using pulse oximeter sensors similar to those currently employed in existing wrist-worn devices. We describe WristO2, which uses pulse oximetry, as well as motion data from an IMU sensor (gyroscope and an accelerometer), to detect and reject unreliable readings. WristO2 consists of a pipeline of

1https://www.samsung.com/global/galaxy/galaxy-s8/specs/  
automated feature extraction and a gradient boosting classifier that labels signals as reliable or unreliable. In a preliminary study, we show that WristO2 reduces the average error in SpO2 readings from 14.5 percentage points to 1.5 percentage points compared to a baseline implementation, while generating a reading on average at least every three minutes. We also investigate the effects of skin tone, IMU data, and per-user fine-tuning on the accuracy of WristO2.

II. THE CHALLENGE OF WRIST-BASED PULSE OXIMETRY

Pulse oximeters allow non-invasive monitoring of blood oxygen saturation. Light emitted by LEDs interact with the users blood and is then captured by photodetectors to produce a photoplethysmogram, or PPG. An estimate of oxygen saturation is produced from the PPG by calculating a ratio of ratios between the amount of red (660nm, absorbed mostly by non-oxygenated blood) and infrared (940nm, absorbed mostly by oxygenated blood) light detected, as described in Equation 1:

\[ \text{SpO}_2 = y_0 - m \times \left( \frac{AC_{\text{Red}}/DC_{\text{Red}}}{AC_{\text{IR}}/DC_{\text{IR}}} \right) \] (1)

AC and DC denote the alternating and direct current measured by the photodetector for each light source. A small window of the PPG data, usually four seconds, is used to calculate an SpO2 reading [10]. In general, transmissive pulse oximeters that cover the fingertip are considered more accurate than reflective ones: motion and ambient light artifacts can produce unreliable SpO2 readings, especially when measured from the wrist, where good contact is not guaranteed.

To evaluate the feasibility of using pulse oximetry to measure SpO2 from the wrist, we collect data using a MAX30102 sensor from Maxim Integrated [11], a manufacturer-grade reflective pulse oximeter similar to those found in modern commodity smartwatches, fitness bands, phones, and other personal electronics. We use the SpO2 calculation algorithm provided by the manufacturer as a baseline. We also consider an enhanced version of the baseline algorithm that discards PPG readings from 14.5 percentage points to 1.5 percentage points compared to the fingertip readings, which we consider to be too big a margin of error, given that the healthy range for individuals is 90% to 100%.

Figure 2 shows two PPG traces obtained from the same user at the same time using identical reflective sensors. The top PPG was captured from a fingertip-worn reflective pulse oximeter over several seconds. The strong periodic signal captures the change in flow of oxygenated blood through the fingertip. The bottom PPG was taken from the wrist. Even with clean contact to the skin, this trace is much noisier. Algorithms used to produce reliable SpO2 readings from a wrist-worn sensor must be able to mitigate these errors.

III. SYSTEM

WristO2 is a filtering system designed to identify which signal windows captured from the wrist-worn sensor will produce reliable SpO2 readings. WristO2 uses statistical machine learning techniques to train a model to classify wrist-readings as reliable or unreliable. As input to our classifier we compute features from 4 signal sources: red and infrared LEDs, and gyroscope and accelerometer magnitude.

To collect training and evaluation data, we take PPG measurements from two points of contact on a single user, namely the fingertip and the wrist. PPG traces and IMU data are collected from all sensors simultaneously. A custom wrist-worn device is attached to the dominant hand of a user during experiments, and allows for users to maintain range of motions in their wrist. Movement throughout the duration of experiments is encouraged. Custom hardware was required because commodity wrist-worn PPG sensors either did not have the appropriate LEDs (Samsung), or did not provide...
access to raw sensor data at sub-second granularity (Apple). Readings are collected and synchronized from sensors using a microcontroller at a rate of 25Hz.

The wrist-worn device in Figure 4 houses a MAX30102 and an MPU9250 IMU sensor, sewn into a fitness band for stability and consistency across measurements. The user wears the device with the pulse oximeter facing the top of the wrist so that it matches the sensor placement in a vast majority of consumer grade wristbands and smartwatches.

We use a signal window size of 100 sensor readings, or 4 seconds of data, when extracting features. We use the level of agreement with a more reliably collected signal as the ground truth label. Specifically, agreement between the same MAX30102 sensor applied to both the wrist and fingertip. We label wrist-worn readings as reliable if they are within a margin of error from the fingertip reading. Initially this threshold is set to ±2.0 percentage points. The readings from the fingertip sensor are only used for creating the reliability label during training, and not used at test time.

A. Reliability Classifier

To train our classifier, we first extract features from the wrist-worn sensors to be used as inputs when predicting the reliability of the signal. We sample 4 signals, two raw LED radiance signals from the red and infrared channels, and two signals from the magnitude of the gyroscope and accelerometer in the MPU9250.

We use the Tsfresh [13] library to extract features from the time series data. The library calculates and evaluates significance for a comprehensive list of features. Depending on the training data provided, 900-1000 features are selected by the library, though many have a very low significance and can be removed without affecting classifier performance. Table I show the 14 most significant features. Features calculated on the red and infrared channels provide the most significance, followed by the gyroscope. Accelerometer data provides significantly less information.

The classifier is trained with XGBoost [14] using a learning rate of 0.1, 100 estimators, a max depth of 3, minimum child weight of 3, regularization alpha of 0.3, a subsample ratio of 0.9, and a logistic binary objective. Classifier selection and tuning was performed empirically using a cross-validated hyperparameter search. The classifier is trained using leave-one-out cross validation across participants, and evaluated on held out validation sets. During training, non-overlapping windows are used to ensure that feature data is independent, however new data is evaluated with a sliding window. Data is split in time to ensure a look-ahead bias is avoided.

B. Data Collection

We collect data from 10 participants. Each user has the wrist-band with the pulse oximeter and IMU sensor attached to their dominant hand, and a MAX30102 sensor attached directly to their fingertip on the opposing hand. Trials on each participant are conducted for approximately 12 minutes, during which time users are encouraged to continue using their dominant hand in an effort to provide the most naturally acquired readings. To reduce motion artifacts when acquiring ground truth readings, participants are asked to keep their non-dominant hand motionless for the duration of the experiment. Users range from 20-55 years of age and vary in skin colour. The proportion of unreliable readings measured from each user ranges from 30%-99%, relative to the fingertip ground truth. Variables such as skin colour, device tightness, wrist thickness, movement, and ambient light, can all affect the number of reliable readings.

IV. Evaluation

We evaluate WristO₂ on three metrics. First, precision, the fraction of truly reliable readings out of those classified as reliable. We emphasize precision over recall: few intermittent reliable SpO₂ readings are preferable to a continuous stream of potentially false readings. Intuitively, due to the relatively low fluctuations of oxygen saturation levels, SpO₂ can be reliably interpolated with frequent enough measures. Second, we report the root mean squared error, RMSE, of readings taken from the wrist-worn sensor as compared with the fingertip sensor. We take the RMSE before pruning values with WristO₂, and then calculate the RMSE after pruning to determine any improvement. Finally, to avoid pruning too many PPG traces and making the system unusable, we measure the time between valid readings: the longest interval where WristO₂ produces no SpO₂ readings.

TABLE I

<table>
<thead>
<tr>
<th>Tsfresh function call for top 14 features.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. longest_strike_below_mean('ir')</td>
</tr>
<tr>
<td>2. autocorrelation('ir', 6)</td>
</tr>
<tr>
<td>3. autocorrelation('ir', 5)</td>
</tr>
<tr>
<td>4. autocorrelation('ir', 7)</td>
</tr>
<tr>
<td>5. autocorrelation('ir', 8)</td>
</tr>
<tr>
<td>6. autocorrelation('ir', 9)</td>
</tr>
<tr>
<td>7. cid_ce('ir', normalize=True)</td>
</tr>
<tr>
<td>8. autocorrelation('ir', 4)</td>
</tr>
<tr>
<td>9. ar_coefficient('red', {'coeff': 0, 'k': 10})</td>
</tr>
<tr>
<td>10. splt_welch_density('red', {'coeff': 2})</td>
</tr>
<tr>
<td>11. ar_coefficient('ir', {'coeff': 0, 'k': 10})</td>
</tr>
<tr>
<td>12. mean('gyro')</td>
</tr>
<tr>
<td>13. sum_value('gyro')</td>
</tr>
<tr>
<td>14. fft_coefficient('gyro', {'coeff': 0, 'attr': 'abs'})</td>
</tr>
</tbody>
</table>
A. Performance of WristO2

Filtering readings with WristO2 shows a drastic reduction in error compared to existing methods. Table II shows the mean RMSE across all users for the baseline algorithm, enhanced algorithm and WristO2; Figure 5 shows the resulting absolute errors of wrist sensor readings. WristO2 achieved an average precision of 72% across users. WristO2 reduces RMSE of SpO2 measurement by an order of magnitude compared to the baseline algorithm, and by 4.5 times for the enhanced baseline.

To illustrate the effect of WristO2, Figure 6 shows part of a trace for a single user. The blue line representing the enhanced algorithm applied to a PPG trace collected from the wrist, and the orange line representing the enhanced algorithm applied to the signal collected from the fingertip during the same session. Finally, the black line represents the readings remaining after WristO2 prunes unreliable results. Spikes and inaccuracies are clearly visible even with the Enhanced algorithm. WristO2 successfully rejects many of these unreliable readings.

The reduction in error comes at the cost of producing less readings compared to the existing algorithms. Figure 7 shows the CDF of the maximum size of an interval where WristO2 produced no readings across all users. The average interval between readings across all users is approximately 3 minutes, with the worst case for a single user at approximately 6 minutes and 40 seconds. Given that the current state-of-the-art for acquiring reliable SpO2 readings requires a user to actively clip a commercial pulse oximeter to their fingertip and wait, the time between readings from existing methods would be collected in the order of several hours or even half a day. A mean interval of less than 3 minutes for automatic collection of reliable readings is a dramatic improvement.

B. Effect of Skin Tone

As discussed in section VI, it has been shown that it is more difficult to collect a reliable signal when darker pigment exists on the skin, whether naturally or artificially from tattoo ink. We aim to quantify potential difficulty in collecting reliable PPG traces from users of various skin tones. As five of our ten participants had light skin tone, we partition them into two groups lighter- or darker-skinned, and evaluate classifiers with all permutations of training and testing groups. Mean (and Std. Dev.) are shown across users of the testing group. In cases where the training and testing groups are the same, leave-one-out cross-validation is used across user’s of the group.

Table III shows that precision is improved when classifying on lighter skin as opposed to darker skin, regardless of the skin-tone used during training. We caution that our sample size is too small to draw strong conclusions about the magnitude of effects, and much more data will be needed to adequately characterize performance discrepancies between pigment groups. Regardless, in both groups the error is reduced by WristO2; and we have shown that the classifier will generalize to pigment colours that it was not trained on.

C. Per-User Training

We explore the viability of building a personalized WristO2 classifier on a per user basis. Consider a user that has a wrist-worn device with a pulse oximeter capable of measuring SpO2, such as a smartwatch, and a similar fingertip sensor such as those that exist in the back of certain Samsung smartphones. During a calibration phase, the user can be instructed to wear the smart watch while simultaneously pressing their finger against the sensor on the smart phone. Once sufficient calibration data can be captured, the classifier can be retrained with the additional data to provide the user with more reliable readings from smartwatch.

We train a classifier with 9 users and test it on an unseen user. We then add 2 minutes of data from the previously unseen user to the training set and retrain the model. And again with 10 minutes. The results are summarized in Table IV. Pruning signal windows with WristO2 reduces the RMSE to 3.8% even when no calibration data is used. Using a small amount of user.

### Table II

<table>
<thead>
<tr>
<th>Skin tone</th>
<th>Precison</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>Testing</td>
<td>WristO2</td>
</tr>
<tr>
<td>Dark</td>
<td>Dark</td>
<td>37% (28%)</td>
</tr>
<tr>
<td>Dark</td>
<td>Light</td>
<td>80% (20%)</td>
</tr>
<tr>
<td>Light</td>
<td>Dark</td>
<td>42% (32%)</td>
</tr>
<tr>
<td>Light</td>
<td>Light</td>
<td>69% (27%)</td>
</tr>
</tbody>
</table>

### Table III

<table>
<thead>
<tr>
<th>Skin tone</th>
<th>Effect of Skin Tone on SpO2 Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Light</td>
<td>Dark</td>
</tr>
<tr>
<td>Dark</td>
<td>Light</td>
</tr>
<tr>
<td>Dark</td>
<td>Dark</td>
</tr>
</tbody>
</table>

Fig. 5. CDF of RMSE for existing algorithms and WristO2 for all users.

Fig. 6. SpO2 measurements over time from one trace.

Fig. 7. CDF of longest delay between valid readings.
specific training data on top of the original training set further reduces the RMSE by up to 0.7 percentage points.

D. Importance of Accelerometer and Gyroscope

To study the effect of the features extracted from the IMU signal on classification, we evaluated the WristO₂ classifier with different combinations of features from the LEDs and IMU sensor as input. Table V summarizes the results.

Approximately half of the features extracted and selected by the TSfresh pipeline are features from the IMU. Although features from the LED channels alone contribute to a significant reduction in the RMSE, adding the 497 features extracted from the IMU signals further reduces the RMSE to 1.5%. It is sensible that the LED channels contribute a majority of the performance increase considering the LEDs are used directly to calculate \( \text{SpO}_2 \). We verify that this is the case by training the classifier with traces solely from the IMU, which shows a negligible increase in performance.

E. Effects of window size and varied thresholds

Window size is measured to be optimal at the original set value of 100 readings, or 4 seconds. Reducing the window to 50 increases RMSE by approximately 1%, and reducing it to 25 increases RMSE by an additional 1.5%. Smaller window sizes could be used to reduce compute requirements.

Figure 8 shows varied sizes of reliability thresholds used to generate labels, and their corresponding RMSE results.

Although it would appear that lower threshold values improve results overall, it is worth noting that the frequency of acquired readings is inversely proportional to the expected RMSE, as Figure 9 demonstrates for the varied threshold sizes.

V. LIMITATIONS AND FUTURE WORK

Our small sample size warrants a larger study to validate results. Ideally the study should be deployed in a clinical setting where data can be collected from patients with a variety of conditions affecting arterial oxygen saturation. Validating WristO₂ on patients outside of the healthy range is essential to guaranteeing effectiveness. Our results support that more study is needed on the effect of skin tone on wrist-worn pulse oximeters.

We consider extending the classifier to multi-label classification for different confidence ranges, or regression. That is, predict not whether a signal will produce a reliable label within a certain threshold, but rather the confidence that the label will be produced within multiple thresholds (i.e., 1%, 3%, and 10%). A similar technique was used in [15] to control the trade-off between accuracy and time between readings in respiratory rate extraction.

The feature set could be pruned enough such that feature extraction can be performed live on a mobile device. Work similar to Sidewinder [16] could be used to offload signal reliability calculation to a lower powered processor, and subsequently wake the device when a usable signal is detected. Deploying WristO₂ on an existing smartwatch platform that provides low level access to the LED sensor readings and the correct spectrum’s could improve experimental consistency and results through higher quality hardware.

VI. RELATED WORK

Existing work on reflective sensors is focused on heart rate measurement, such as rule based detection of heart rate for reliability [17]. Ra et al. perform reliability detection in the context of wrist heart-rate measurement on existing smartwatches [18]. There has been work done to improve reliability in fingertip sensors through signal preprocessing and noise reduction [10] [19]. Possible wearability sites, including the wrist, and various sensor configurations have been considered in the context of telehealth monitoring [6], [20]. Reflective pulse oximeters are widely used and studied in medicine in
between red and infrared wavelengths to improve SpO2 worn PPG signals, and a brief review of fundamentals.

Jarchi et al. [23] explored using the common information of state-of-the-art research on heart rate estimation from wrist-as infant monitoring [21]. See [22] for a comprehensive review places where transitive pulse oximeters are not feasible, such as pediatric environments. Yan et al. [25] used more sophisticated feature extraction to remove motion from pulse oximeters used for telehealth monitoring.

Severinghaus et al. [26] showed that bias in SpO2 measurements increases during a state of anemia (low red blood cell count). Emery et al. [27] and Cote et al. [28] explore the effects of dark skin pigmentation and ink in fingertip worn pulse oximeters. Lee et al. [29] showed that lower true pulse oximetry values were overestimated for a set of people from Singapore due to darker pigmentation. Sjoding et al. [30] conducted a study of fingertip pulse oximetry in an ICU setting with similar conclusions. Ray et al. [31] find that for dark skin, smartwatches report lower-confidence heart rate measurements even when they are reliable.

Liqiat et al. are currently working on using wrist-worn devices to aid COPD patients in treatment and disease management in the context of the WearCOPD project [32], [33]. Although they currently do not employ SpO2 in their consideration of patient health, reliable SpO2 readings could improve COPD monitoring [2].

VII. CONCLUSION

In this work we study the reliability of SpO2 measurements from a wrist-worn pulse oximeter, and show that existing algorithms often provide unreliable readings. We propose WristO2, which uses automated feature extraction and statistical machine learning to identify reliable peripheral oxygen saturation readings taken from the wrist. After pruning unreliable results with WristO2, we show that error in measurements taken from the wrist can be reduced by an order of magnitude. Additionally we demonstrate that after pruning results, the frequency of reliable readings is still high enough to be useful. We discuss the effects of skin tone, IMU information, and propose platforms for user level calibration.

REFERENCES


