WristO$_2$: 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 (SpO_2) 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 SpO_2. However, current techniques for SpO_2 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 SpO_2 measurements from the wrist. Using a custom wrist-worn pulse oximeter, we find that existing algorithms used in traditional fingertip SpO_2 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 SpO_2 sensors and measurement algorithms.

Next, based on our findings, we propose WristO$_2$, an alternative approach for reliable SpO_2 sensing. By selectively pruning unreliable data, WristO$_2$ 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 (SpO_2) 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 SpO_2 monitoring only provide intermittent readings since they require active user interaction. To measure their SpO_2 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 S8[1]) that requires the user to press a fingertip against the sensor to obtain an SpO_2 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 SpO_2. Additionally, the device is in constant contact with the user's skin which eliminates the need for active interaction. In practice, however, SpO_2 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 SpO_2 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 SpO_2 readings, users have found the measurements to be unreliable[2] 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 SpO_2 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 SpO_2 signal can be taken automatically from the wrist using pulse oximeter sensors similar to those currently employed in existing wrist-worn devices. We describe WristO$_2$, which uses pulse oximetry, as well as motion data from an IMU sensor (gyroscope and an accelerometer), to detect and reject unreliable readings. WristO$_2$ consists of a pipeline of

[1]https://www.samsung.com/global/galaxy/galaxy-s8/specs/

[2]https://www.washingtonpost.com/technology/2020/09/23/apple-watch-oximeter/

automated feature extraction and a gradient boosting classifier that labels signals as reliable or unreliable. In a preliminary study, we show that WristO$_2$ reduces the average error in SpO$_2$ 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 WristO$_2$.

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:

$$SpO_2 = y_0 - m \times \left(\frac{AC_{Red}/DC_{Red}}{AC_{IR}/DC_{IR}} \right) \quad (1)$$

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 SpO$_2$ 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 SpO$_2$ readings, especially when measured from the wrist, where good contact is not guaranteed.

To evaluate the feasibility of using pulse oximetry to measure SpO$_2$ 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 SpO$_2$ calculation algorithm provided by the manufacturer as a baseline. We also consider an *enhanced* version of the baseline algorithm that discards PPG readings where the Pearson correlation coefficient between the red and infrared wavelengths is below 0.4, a potential indicator of motion artifacts.

We compare the SpO$_2$ output of the MAX30102 sensor from each algorithm with a Berry BM3000B [12], a commercial transmissive oximeter. We collected data from 10 subjects who wore the two SpO$_2$ readers on the non-dominant hand (reflective on the index finger, transmissive on the middle finger) for a period of 12 minutes each. The mean measurement difference between devices is 1.84%, with standard deviation of 1.32% – indicating good agreement between sensors except a small bias, which remained consistent across all users.

After recalibrating the reflective sensor to remove bias, the mean absolute difference drops to 1.01% with standard deviation of 0.77% indicating strong agreement. Over 99% of reflective sensor readings are within $\pm 2\%$ of the transmissive sensor readings. We conclude that our reflective SpO_2 sensor produces reliable measurements when placed on the fingertip. In the remainder of this paper, we use measurements collected

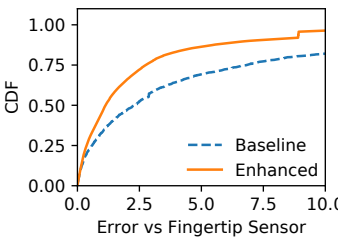

Fig. 1. CDF of absolute difference between wrist and fingertip readings.

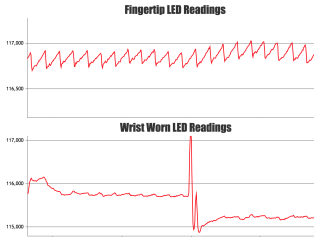

Fig. 2. PPG trace for a fingertip vs. wrist attached sensor (taken from the PPG web platform described in Section III).

with our reflective SpO_2 sensor mounted on the fingertip as ground truth. Using the same sensor allows us to measure reliability across different measurement sites, rather than across different hardware manufacturers.

Naïvly applying existing algorithms to PPG traces obtained from the wrist results in mostly unreliable SpO$_2$ measurements. Figure 1 shows the cumulative distribution function (CDF) of absolute error of readings taken from the wrist using both existing algorithms compared with the same MAX30102 sensor on the fingertip. Despite the increase in performance of the enhanced algorithm, more than 10% of the readings across all users have an error of 5 percentage points or more 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 SpO_2 readings from a wrist-worn sensor must be able to mitigate these errors.

III. System

WristO$_2$ is a filtering system designed to identify which signal windows captured from the wrist-worn sensor will produce reliable SpO$_2$ readings. WristO$_2$ 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

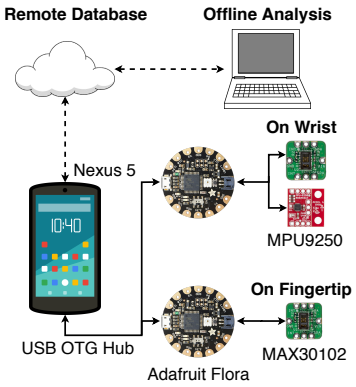

Fig. 3. The data collection platform.

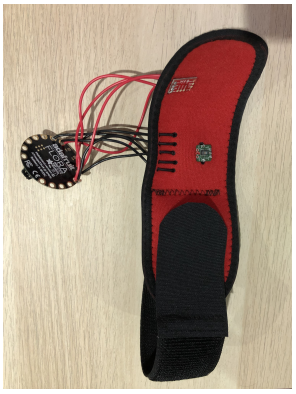

Fig. 4. Custom wrist wearable.

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 WristO_2 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 SpO_2 readings are preferable to a continuous stream of potentially false readings. Intuitively, due to the relatively low fluctuations of oxygen saturation levels, SpO_2 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 WristO_2, 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 WristO_2 produces no SpO_2 readings.

TABLE I
TSFRESH FUNCTION CALL FOR TOP 14 FEATURES.

1. longest_strike_below_mean('ir')	8. autocorrelation('ir', 4)
2. autocorrelation('ir', 6)	9. ar_coefficient('red', {"coeff": 0, "k": 10})
3. autocorrelation('ir', 5)	10. spkt_welch_density('red', {"coeff": 2})
4. autocorrelation('ir', 7)	11. ar_coefficient('ir', {"coeff": 0, "k": 10})
5. autocorrelation('ir', 8)	12. mean('gyro')
6. autocorrelation('ir', 9)	13. sum_values('gyro')
7. cid_ce('ir', normalize=True)	14. fft_coefficient('gyro', {"coeff": 0, "attr": "abs"})

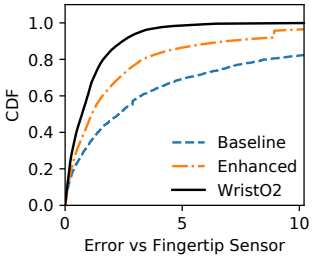

Fig. 5. CDF of RMSE for existing algorithms and $WristO_2$ for all users.

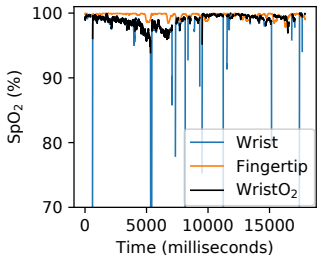

Fig. 6. SpO_2 measurements over time from one trace.

TABLE II
MEAN RMSE (AND STANDARD DEVIATION) ACROSS ALL 10 USERS.

Baseline RMSE	Enhanced RMSE	$WristO_2$ RMSE
14.5% (6.9%)	6.7% (4.4%)	1.5% (0.7%)

A. Performance of $WristO_2$

Filtering readings with $WristO_2$ 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 $WristO_2$; Figure 5 shows the resulting absolute errors of wrist sensor readings. $WristO_2$ achieved an average precision of 72% across users. $WristO_2$ reduces RMSE of SpO_2 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 $WristO_2$, 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 $WristO_2$ prunes unreliable results. Spikes and inaccuracies are clearly visible even with the Enhanced algorithm. $WristO_2$ 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 $WristO_2$ 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 SpO_2 readings requires a user to actively

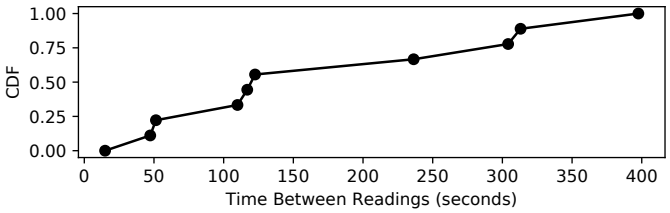

Fig. 7. CDF of longest delay between valid readings.

TABLE III
EFFECTS OF SKIN TONE ON SpO_2 ACCURACY.

Skin tone		Precision	RMSE	
Training	Testing	$WristO_2$	Enhanced	$WristO_2$
Dark	Dark	37% (28%)	8.0% (5.0%)	1.6% (0.5%)
Dark	Light	80% (20%)	5.2% (3.2%)	1.3% (1.0%)
Light	Dark	42% (32%)	8.0% (5.0%)	4.3% (3.3%)
Light	Light	69% (27%)	5.2% (3.2%)	2.6% (2.0%)

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 $WristO_2$; 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 $WristO_2$ classifier on a per user basis. Consider a user that has a wrist-worn device with a pulse oximeter capable of measuring SpO_2, 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 $WristO_2$ reduces the RMSE to 3.8% even when no calibration data is used. Using a small amount of user

TABLE IV
PERFORMANCE WITH USER CALIBRATION DATA.

	Precision	RMSE	
Calibration Data	**WristO$_2$**	**Enhanced**	**WristO$_2$**
None	33%	9.3%	3.8%
+ 2 minutes	34%	9.3%	3.3%
+ 10 minutes	41%	9.3%	3.1%

TABLE V
EFFECTS OF IMU FEATURES ON CLASSIFICATION.

		Precision	RMSE	
Input	**# Features**	**WristO$_2$**	**Enhanced**	**WristO$_2$**
LED Only	489	69% (19%)	6.7% (4.4%)	1.8% (0.9%)
IMU Only	497	47% (35%)	6.7% (4.4%)	5.5% (5.2%)
LED + IMU	986	73% (19%)	6.7% (4.4%)	1.5% (0.7%)

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 WristO$_2$ 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 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.

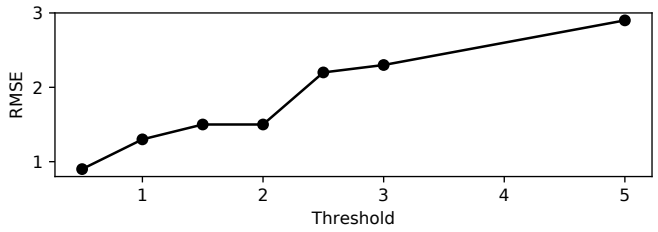

Fig. 8. RMSE for different thresholds.

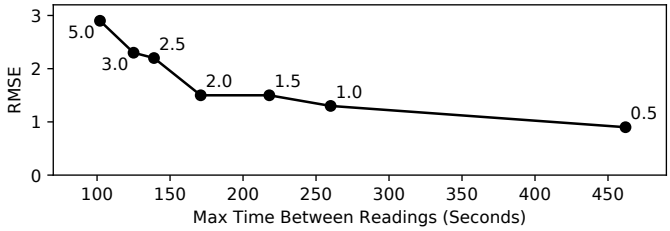

Fig. 9. The trade-off between quantity and quality of readings for different reliability thresholds. For every threshold, the X axis shows the resulting worst-case interval between reliable readings, and the Y axis shows the resulting RMSE.

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 WristO$_2$ 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 WristO$_2$ 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

places where transitive pulse oximeters are not feasible, such as infant monitoring [21]. See [22] for a comprehensive review of state-of-the-art research on heart rate estimation from wrist-worn PPG signals, and a brief review of fundamentals.

Jarchi et al. [23] explored using the common information between red and infrared wavelengths to improve SpO_2 measurements from the wrist, although their study is based on only 5 subjects. Our approach is complementary, and can be used in parallel. Yao et al. [24] used simple motion sensing to remove noise from movement artifacts to improve signal reliability in ambulatory 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 SpO_2 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 oximetery 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.

Liaqat 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 SpO_2 in their consideration of patient health, reliable SpO_2 readings could improve COPD monitoring [2].

VII. Conclusion

In this work we study the reliability of SpO_2 measurements from a wrist-worn pulse oximeter, and show that existing algorithms often provide unreliable readings. We propose $WristO_2$, 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 $WristO_2$, 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

[1] R. T. Brouillette, A. Morielli, A. Leimanis, K. A. Waters, R. Luciano, and F. M. Ducharme, "Nocturnal pulse oximetry as an abbreviated testing modality for pediatric obstructive sleep apnea," *Pediatrics*, vol. 105, no. 2, pp. 405–412, 2000.

[2] P. Sliwinski, M. Lagosz, D. Gorecka, and J. Zielinski, "The adequacy of oxygenation in COPD patients undergoing long-term oxygen therapy assessed by pulse oximetry at home," *Eur. Respir. J.*, vol. 7, no. 2, pp. 274–278, 1994.

[3] A. H. Taenzer, J. B. Pyke, S. P. McGrath, and G. T. Blike, "Impact of pulse oximetry surveillance on rescue events and intensive care unit transfersa before-and-after concurrence study," *Anesthesiology*, vol. 112, no. 2, pp. 282–287, 2010.

[4] S. Shah, K. Majmudar, A. Stein, N. Gupta, S. Suppes, M. Karamanis, J. Capannari, S. Sethi, and C. Patte, "Novel use of home pulse oximetry monitoring in COVID-19 patients discharged from the emergency department identifies need for hospitalization," *Acad. Emerg. Med.*, vol. 27, no. 8, pp. 681–692, 2020.

[5] A. J. Enoch, M. English, and S. Shepperd, "Does pulse oximeter use impact health outcomes? a systematic review," *Archives of disease in childhood*, vol. 101, no. 8, pp. 694–700, 2016.

[6] Y. Mendelson and C. Pujary, "Measurement site and photodetector size considerations in optimizing power consumption of a wearable reflectance pulse oximeter," in *EMBC*, vol. 4, 2003, pp. 3016–3019.

[7] A. Shcherbina, C. M. Mattsson, D. Waggott, H. Salisbury, J. W. Christle, T. Hastie, M. T. Wheeler, and E. A. Ashley, "Accuracy in wrist-worn, sensor-based measurements of heart rate and energy expenditure in a diverse cohort," *J. Pers. Med.*, vol. 7, no. 2, p. 3, 2017.

[8] B. Bent, B. A. Goldstein, W. A. Kibbe, and J. P. Dunn, "Investigating sources of inaccuracy in wearable optical heart rate sensors," *NPJ Digit. Med.*, vol. 3, no. 1, pp. 1–9, 2020.

[9] S. Preejith, A. S. Ravindran, R. Hajare, J. Joseph, and M. Sivaprakasam, "A wrist worn spo 2 monitor with custom finger probe for motion artifact removal," in *EMBC*, 2016, pp. 5777–5780.

[10] P. M. Mohan, A. A. Nisha, V. Nagarajan, and E. S. J. Jothi, "Measurement of arterial oxygen saturation (spo 2) using ppg optical sensor," in *ICCSP*, 2016, pp. 1136–1140.

[11] M. Integrated, "MAX30102 high-sensitivity pulse oximeter and heart-rate sensor for wearable health," 2018. [Online]. Available: https://www.maximintegrated.com/en/products/sensors-and-sensor-interface/MAX30102.html/tb_tab0

[12] Berry Medical, "BM3000B usb pulse meter," http://www.shberrymed.com/usb-pulse-meter-bm3000b-p00037p1.html.

[13] M. Christ, N. Braun, J. Neuffer, and A. W. Kempa-Liehr, "Time series feature extraction on basis of scalable hypothesis tests (tsfresh–a Python package)," *Neurocomputing*, 2018.

[14] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *KDD*, 2016, pp. 785–794.

[15] D. Liaqat, M. Abdalla, P. Abed-Esfahani, M. Gabel, T. Son, R. Wu, A. Gershon, F. Rudzicz, and E. D. Lara, "WearBreathing: Real world respiratory rate monitoring using smartwatches," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 3, no. 2, Jun. 2019.

[16] D. Liaqat, S. Jingoi, E. de Lara, A. Goel, W. To, K. Lee, I. De Moraes Garcia, and M. Saldana, "Sidewinder: An energy efficient and developer friendly heterogeneous architecture for continuous mobile sensing," *ACM SIGARCH*, vol. 44, no. 2, pp. 205–215, 2016.

[17] A. Al Ali, D. S. Breed, J. J. Novak, and M. E. Kiani, "Pulse oximetry data confidence indicator," Jan. 27 2004, US Patent 6,684,090.

[18] H.-K. Ra, J. Ahn, H. J. Yoon, D. Yoon, S. H. Son, and J. Ko, "I am a "smart" watch, smart enough to know the accuracy of my own heart rate sensor," in *HotMobile*, 2017, pp. 49–54.

[19] J. Yao and S. Warren, "A short study to assess the potential of independent component analysis for motion artifact separation in wearable pulse oximeter signals," in *EMBC*, 2005, pp. 3585–3588.

[20] Y. Mendelson, R. Duckworth, and G. Comtois, "A wearable reflectance pulse oximeter for remote physiological monitoring," in *EMBC*, 2006, pp. 912–915.

[21] D. R. Tobler, M. K. Diab, and R. J. Kopotic, "Fetal pulse oximetry sensor," Sep. 4 2001, US Patent 6,285,896.

[22] D. Biswas, N. Simões-Capela, C. Van Hoof, and N. Van Helleputte, "Heart rate estimation from wrist-worn photoplethysmography: A review," *IEEE Sensors Journal*, vol. 19, no. 16, pp. 6560–6570, 2019.

[23] D. Jarchi, D. Salvi, C. Velardo, A. Mahdi, L. Tarassenko, and D. A. Clifton, "Estimation of HRV and SpO2 from wrist-worn commercial sensors for clinical settings," in *BSN*, 2018, pp. 144–147.

[24] J. Yao and S. Warren, "A novel algorithm to separate motion artifacts from photoplethysmographic signals obtained with a reflectance pulse oximeter," in *EMBC*, vol. 1, 2004, pp. 2153–2156.

[25] Y.-S. Yan and Y.-T. Zhang, "An efficient motion-resistant method for wearable pulse oximeter," *IEEE Trans. Inf. Technol. Biomed.*, vol. 12, no. 3, pp. 399–405, 2008.

[26] J. W. Severinghaus and S. O. Koh, "Effect of anemia on pulse oximeter accuracy at low saturation," *J. Clin. Monit.*, vol. 6, no. 2, pp. 85–88, 1990.

[27] J. Emery, "Skin pigmentation as an influence on the accuracy of pulse oximetry," *J. Perinatol.*, vol. 7, no. 4, pp. 329–330, 1987.

[28] C. J. Coté, E. A. Goldstein, W. H. Fuchsman, and D. C. Hoaglin, "The effect of nail polish on pulse oximetry," *Anesthesia and analgesia*, vol. 67, no. 7, pp. 683–686, 1988.

[29] K. Lee, K. Hui, W. Tan, and T. Lim, "Factors influencing pulse oximetry as compared to functional arterial saturation in multi-ethnic Singapore," *Singapore Medical Journal*, vol. 34, pp. 385–385, 1993.

[30] M. W. Sjoding, R. P. Dickson, T. J. Iwashyna, S. E. Gay, and T. S. Valley, "Racial bias in pulse oximetry measurement," *New Eng. J. Med.*, vol. 383, no. 25, pp. 2477–2478, 2020.

[31] I. Ray, D. Liaqat, M. Gabel, and E. de Lara, "Skin tone, confidence, and data quality of heart rate sensing in WearOS smartwatches," in *PerCom Workshops*, 2021.

[32] D. Liaqat, I. Thukral, P. Sin, H. Alshaer, F. Rudzicz, E. de Lara, R. Wu, and A. Gershon, "Poster: Wearcopd - monitoring COPD patients remotely using smartwatches," in *MobiSys*, 2016, pp. 139–139.

[33] R. Wu, D. Liaqat, E. de Lara, T. Son, F. Rudzicz, H. Alshaer, P. Abed-Esfahani, and A. Gershon, "Feasibility of using a smartwatch to intensively monitor patients with chronic obstructive pulmonary disease: Prospective cohort study," *JMIR mHealth and uHealth*, vol. 6, no. 6, p. e10046, 2018.