# WRISTO$_2$: RELIABLE PERIPHERAL OXYGEN SATURATION READINGS FROM WRIST-WORN PULSE OXIMETERS

A PREPRINT

**Caleb Phillips**
University of Toronto
caleb@cs.toronto.edu

**Daniyal Liaqat**
University of Toronto and Vector Institute
dliaqat@cs.toronto.edu

**Moshe Gabel**
University of Toronto
mgabel@cs.toronto.edu

**Eyal de Lara**
University of Toronto
delara@cs.toronto.edu

June 19, 2019

**ABSTRACT**

Peripheral blood oxygen saturation ($SpO_2$) is a vital measure in healthcare. Modern off-the-shelf wrist-worn devices, such as the Apple Watch, FitBit, and Samsung Gear, have an onboard sensor called a pulse oximeter. While pulse oximeters are capable of measuring both $SpO_2$ and heart rate, current wrist-worn devices use them only to determine heart rate, as $SpO_2$ measurements collected from the wrist are believed to be inaccurate. Enabling oxygen saturation monitoring on wearable devices would make these devices tremendously more useful for health monitoring and open up new avenues of research.

To the best of our knowledge, we present the first study of the reliability of $SpO_2$ sensing from the wrist. Using a custom-built wrist-worn pulse oximeter, we find that existing algorithms designed for fingertip sensing are a poor match for this setting, and can lead to over $90\%$ of readings being inaccurate and unusable. We further show that sensor placement and skin tone have a substantial effect on the measurement error, and must be considered when designing wrist-worn $SpO_2$ sensors and measurement algorithms.

Based on our findings, we propose WristO$_2$, an alternative approach for reliable $SpO_2$ sensing. By selectively pruning 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.

***K*eywords** Health monitoring · Wearable computing · Applied machine learning · Health sensors

## 1 Introduction

Peripheral oxygen saturation ($SpO_2$) has many uses in healthcare monitoring and is a primary vital sign used by nurses and physicians to monitor patients. It is a measure of the amount of oxygenated blood (expressed as a percentage) and it's 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]. However, most methods of $SpO_2$ monitoring are intermittent and require active user interaction. For example, in hospitals, nurses often record $SpO_2$ by attaching a fingertip device to patients during their rounds. At home, individuals concerned about their $SpO_2$ level can purchase similar commercial devices and record their $SpO_2$ a few times per day. As part of the growing mobile health monitoring movement, cell phone manufacturers have recently provided an onboard pulse oximeter on the back of smartphones (e.g Samsung Galaxy S8[1]) that require the user to press a fingertip against the sensor to obtain an $SpO_2$ reading. These

devices all make use of a sensor called a pulse oximeter, which works by emitting light and measuring how much of the light is absorbed by the user's blood.

Adding $SpO_2$ measuring capabilities to wrist-worn devices seems like a logical next step. A significant advantage of monitoring $SpO_2$ on the wrist is that because the device would be in constant contact with the user's skin, it eliminates the need for active interaction from the user and consequently, allows more frequent measurements. The ability to more frequency monitor oxygen saturation levels could provide a useful diagnostic tool, allowing for the development of early interventions that could drastically improve health outcomes and reduce health care costs.

Interestingly, wrist-worn devices such as the Apple Watch, FitBit, and Samsung Gear already contain pulse oximeters. However they only use the data from the pulse oximeter to derive heart rate and not oxygen saturation. The pulse oximeters on these devices are fundamentally the same as the ones used in hospital and commercial fingertip $SpO_2$ monitors, however calculating oxygen saturation from a wrist-worn sensor leads to mostly inaccurate and unreliable [4] data. This is primarily an issue of a poorly fitting devices, wrist/arm movement, low blood perfusion, and interference from ambient light.

Despite the fact that most pulse oximeter readings from a wrist-worn device are unreliable, it is our hypothesis that occasionally readings taken from such a device will be sufficiently reliable. Even if a small fraction of oxygen saturation readings are reliable, as long as they can be confidently identified among a majority of noisy readings, we believe that we can improve the current state of personal oxygen saturation monitoring. 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 a single reliable $SpO_2$ reading every ten minutes, we have succeeded in increasing the amount of available data by almost two orders of magnitude. We have also removed the need for active user interaction.

In this work, we demonstrate that an intermittent reliable $SpO_2$ signal can be taken automatically from the wrist using sensors similar to those currently employed in existing wrist-worn devices, such as the Apple Watch, FitBit, and Samsung Gear (given these devices employed a proper LED configuration). We develop a custom wrist-worn sensor collection platform and record data from ten participants. We implement a system, which we call Wrist$O_2$, that consists of a pipeline of automated feature extraction and a gradient boosting classifier to label signals as reliable or unreliable. Wrist$O_2$ uses pulse oximeter and motion data to detect and reject unreliable data, which reduces the average error from 14.5% to 1.5% compared to a baseline implementation while generating a reading on average at least every three minutes. We also measure the effect of sensor placement and skin tone and show that Wrist$O_2$ is robust to variations in skin tone. Furthermore, we show that Wrist$O_2$ generalizes to unseen skin tones and participants and explore whether training participant specific models is beneficial.

The rest of this paper is organized as follows. Section 2 provides background on pulse oximeters, and shows why $SpO_2$ measurement from the wrist is challenging. In Section 3 describes our approach for building reliable wrist-worn pulse oximeters, and Section 4 details experimental setup and our data collection. In Section 5 we evaluate our approach and compare it to current algorithms. In Section 6 we discuss practical deployment considerations. Section 7 reviews related work, and Section 8 summarizes.

## 2 Background and Motivation

Pulse oximeters are small sensors that allow non-invasive monitoring of heart-rate, blood oxygen saturation, and other health related metrics [5]. A pulse oximeter consists of one or more light emitting diodes (LEDs, usually red and infrared), and a photodetector. Light emitted by the LED interacts with the users blood and is then captured by the photodetector. Because oxygenated hemoglobin and non-oxygenated hemoglobin absorb different wavelengths of light, the amount of each wavelength of light captured by the photodetector indicates the level of oxygen in the blood. This signal (called a *photoplethysmogram*, or PPG), can be used to estimate heart rate, $SpO_2$ and other metrics.

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:

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

*AC* and *DC* denote the alternating and direct current measured by the photodetector for each light source. These terms arise from the periodic nature of the cardiac cycle and the fact that the level of oxygenated arterial blood fluctuates with

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

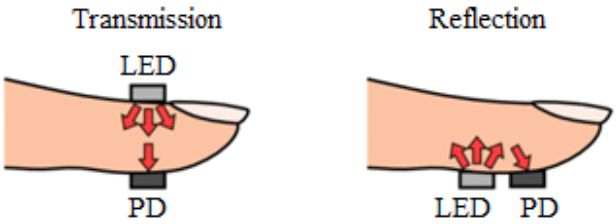


Figure 1: Photodetector and LED placement for transmissive and reflective pulse oximeters (from [7]).

the cardiac cycle whereas the level of non-oxygenated blood in the veins stays fairly constant. The DC component reflects the periodic oxygenated blood while the AC component reflects the constant non-oxygenated blood. This, coupled with the fact that infrared is mostly absorbed by oxygenated blood means taking the ratio of ratios isolates the proportion of oxygenated blood in the artery.

The currents for each reading are taken from a window of a fixed size, 4 seconds (100 samples at $25Hz$) in most cases. The $y_0$ and $m$ terms represent a linear fit for calibration, and would generally be provided by the manufacturer of the specific sensor after they have calibrated the sensor against a ground truth. Equation 1 is based on Beer-Lamberts law, and described in more detail in [6].

### 2.1 Oxygen Saturation Extraction Algorithms

Currently there are two implementations of the $SpO_2$ algorithm, described by Equation 1. The first is a basic implementation that is supplied by the manufacturer for testing purposes. This algorithm simply calculates the ratio of ratios without any filtering of data. We use this algorithm as a baseline for comparing other algorithms and our own implementation.

The second algorithm is an enhancement of the baseline algorithm that implements the same calculation of $SpO_2$, however it attempts to correct for the fact that the signal measured by the photodetector can be noisy. After performing baseline levelling of each signal, a Pearson correlation is calculated between the incoming red and infrared channels. Because the measurement site is the same the intensity of red and infrared light should be highly correlated. If they aren't correlated, it is most likely due to unwanted noise artifacts. Therefore, any signals that produce a correlation value below 0.4 are discarded. The code and a description of the algorithm is available through [2].

Both aforementioned algorithms were originally implemented for an Arduino. For our analysis, we remove the Arduino relevant code and compile the remaining C code to allow for offline analysis and direct comparison between the two algorithms.

### 2.2 Transmissive and Reflective

Pulse oximeters are available in two types; *transmissive* and *reflective*. These are characterized by the relative location of the LEDs and photodetector as shown in Figure 1.

In *transmissive* sensors the LEDs and photodetector sit across from each other so that when clipped to a user's finger, the light from the LED shines through the finger and into the photodetector. Medical settings, such as hospitals and clinics tend to employ transmissive sensors because they are more accurate. However, because they require the sensor to be clipped to the finger, can impede the wearers use of their hands, become uncomfortable after a few minutes, and are generally not well suited for continuous monitoring where the user requires mobility. Other points of attachment, such as the earlobe, are possible, but comfort and mobility remain an issue.

In *reflective* pulse oximetry [5], the photodetector sits beside the LEDs and measures light reflected off the user's tissue. This allows measurement on a wider range of sites on the body (for example, the forehead) and provides greater mobility. The downside to reflective sensors is that the overall amount of light received by the photodetector is less than in transmissive sensors, which means obtaining reliable data from them is more challenging [5]. Under ideal conditions, reflective sensors can still reliable enough to be used in hospitals, but generally only when comfort is more important, such as in NICUs [8]. In a mobile, wrist-worn device, however, conditions are far from ideal. Factors such as ambient light and motion can significantly degrade the quality of PPG data. While this is true for both reflective and transmissive sensors, because reflective sensors are already receiving a weaker signal, these factors have a much greater effect on reflective sensors.

[2] https://www.instructables.com/id/Pulse-Oximeter-With-Much-Improved-Precision/

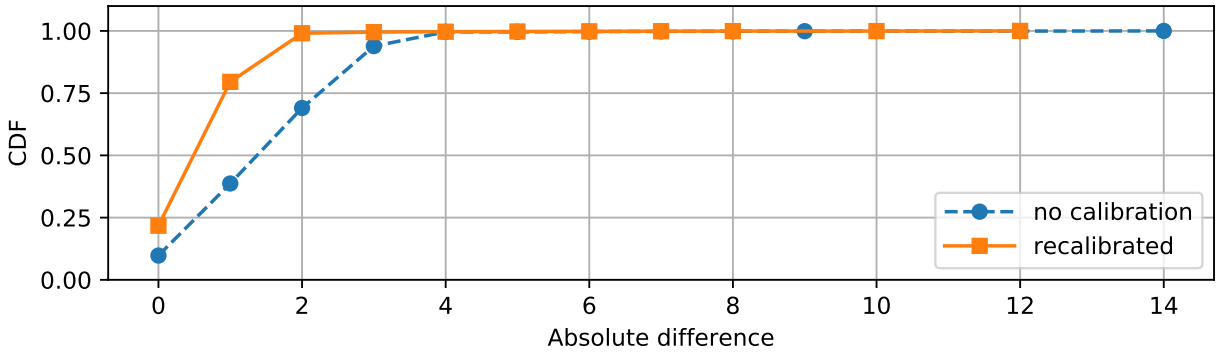


Figure 2: CDF of absolute difference between reflective and transmissive fingerprint sensor readings, before and after recalibration.

Today reflective pulse oximeters are widely available in wrist-worn devices such as smart watches, fitness bands, and cell phones, but given the low accuracy of $SpO_2$ measurement they are generally restricted to measuring heart rate. In fact, major consumer devices have switched to using a single green LED for measurements as it provides better accuracy for heart rate, despite the green LEDs inability to measure $SpO_2$.

### 2.3 Reflective Oximetry from Fingertip

We first show that a reflective $SpO_2$ sensor produces reliable measurements when placed on the fingertip. For this purpose, we compare $SpO_2$ measurements collected with a $SpO_2$ reader we constructed using a MAX30102 reflective sensor (Section 4.1.1), and measurements collected with a Berry BM3000B oximeter [3], a commercial device that uses a transmissive sensor.

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 bias, which remains constant across all users. This bias is most likely a calibration error of one device or the other (the $y_0$ parameter).

We recalibrate the reflective sensor using readings in the first half of each measurement session, resulting in offset of $1.46\%$ over the first half of the data. Figure 2 shows the difference between the recalibrated reflective and transmissive sensors over the second half of the data. 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. Therefore, we conclude that our reflective $SpO_2$ sensor produces reliable measurements when placed on the fingertip. In the reminder of this paper, we use measurements collected with our reflective $SpO_2$ sensor mounted on the fingertip as ground truth.

### 2.4 Reflective Oximetry from Wrist

Unlike reflective fingerprint oximetry which is accurate, a naïvly applying existing methods to PPG traces obtained from the wrist results in unreliable $SpO_2$ measurements.

Figure 3 shows the CDF of absolute error of readings taken from the wrist using both existing algorithms, as compared to an identical fingertip sensor (Section 4 details our on implementation and data collection). 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 given that the healthy range for individuals is 90% to 100%.

Figure 4 shows two PPG traces obtained from the same user at the same time using identical reflective sensors. The left 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 right PPG was taken from the wrist. Even

[3] `http://www.shberrymed.com/usb-pulse-meter-bm3000b-p00037p1.html`

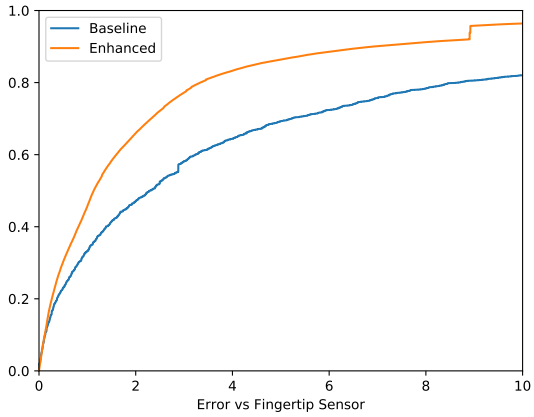


Figure 3: CDF of absolute difference between wrist and fingertip readings.

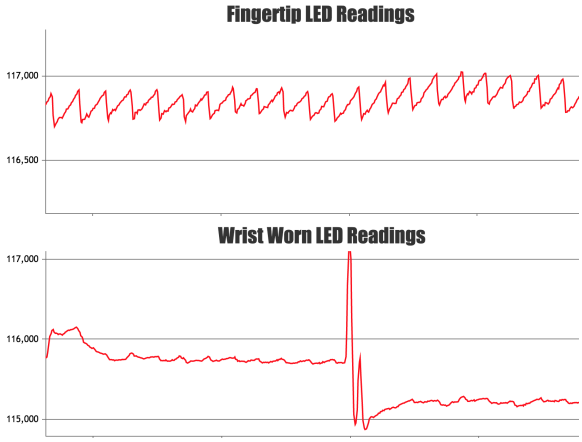


Figure 4: PPG trace for a fingertip vs. wrist attached sensor (taken from the PPG web platform described in Section 4).

with a clean contact with the skin, this PPG is much noisier. The spike in the middle is caused by either motion or ambient light artifact, demonstrating how poor contact with the skin or a user's movements can cause errors and discontinuity in the signal. Algorithms used to produce reliable $SpO_2$ readings from a wrist-worn sensor must be able to mitigate and compensate for these errors.

## 3 Approach

This section outlines Wrist$O_2$, a new approach designed to identify which signal windows captured from the wrist-worn sensor will produce highly reliable $SpO_2$ readings. We still employ the original algorithms for calculating $SpO_2$, however, the goal is to only apply this algorithm to signal windows that will produce a reliable reading.

To identify which signal window will produce a reliable reading, we employ statistical machine learning techniques to train a binary classifier that will classify input data as reliable or unreliable. As input to our classifier, we compute approximately 1000 features Tsfresh[9] from 4 signal sources: red and infrared LEDs, gyroscope magnitude, and accelerometer magnitude. We use a signal window size of 100 sensor readings, or approximately 4 seconds of data when extracting features. This corresponds to the size of the window used to calculate $SpO_2$ by the baseline algorithm. Intuitively, the window size used to calculate the $SpO_2$ will have the greatest effect on it's outcome. To verify this assumption we explore other potential window sizes in section 5.

We will use a level of agreement with a more reliably collected signal as our ground truth. Specifically agreement between the same sensor applied to both the wrist and fingertip. By taking the fingertip readings as truth, we mark wrist-worn readings as reliable if they are within a range of the fingertip readings.

We use various thresholds of agreement between the wrist-worn and fingertip reflective sensors to create the reliability label for classification. Initially for experimentation we set this threshold to $\pm 2.0\%$. That is, if the $SpO_2$ output of the wrist-worn device is within 2 percentage points of the fingertip sensor, we mark the output of the wrist-worn device as reliable. Although we use this threshold in a majority of experiments, we explore the classification results of other reliability threshold values in section 5.

We score the classifier on precision, the ratio of true positive labels over the number of positive instances returned by the classifier, or:

$$Precision = \frac{tp}{tp + fp}$$

The precision of a reliability classifier is the ability of the classifier to only return with a positive score on a reliable result, and minimize the number of false positives. Although this will not produce reliable readings as frequently, it is more desirable for an $SpO_2$ measurement device to provide few intermittent reliable results, rather than a continuous stream of potentially false readings. Intuitively, due to the relatively low fluctuations of true oxygen saturation measurements, $SpO_2$ levels can be reliably interpolated with frequent enough measures. Therefore, we prefer a high true positive score, and a low false positive, with little concern for false negatives.

Considering precision allows us to quantify success of our classifier, we are ultimately concerned with reducing the error in calculated $SpO_2$ readings. Therefore, for a second metric we use the room 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 Wrist$O_2$, and then calculate the RMSE after pruning to determine any improvement.

Because we can potentially remove a bulk of readings while pruning, we add a final metric described as the *time between valid readings*. This measure describes the longest window of silence where Wrist$O_2$ produces no reliable signal with which to calculate $SpO_2$. Although it is desirable for Wrist$O_2$ to reduce our RMSE to zero, we do not want to prune signals so aggressively that we are left with readings that are too infrequent.

Section 4 will describe the experimental setup used to accomplish these requirements, and section 5 will quantify the results of our trained classifiers.

# 4 Implementation

Our work includes a hardware platform for collecting sensor data and a software platform for analyzing that data.

## 4.1 Hardware

Although it would have been desirable to utilize an existing consumer grade device to analyze the current state of wrist-worn pulse oximeters, we encountered two major issues when attempting to select one. The first issue is that the unreliability of $SpO_2$ measurements taken from a wrist-worn pulse oximeter have led manufacturers to focus the technology solely on measuring a users heart rate. Most manufacturers only install a single LED, since heart-rate measurement algorithms implement peak detection and only rely on a single PPG trace. The second issue is that manufacturer APIs are too limited, and do not provide LED reflectance level needed for the PPG trace. In the devices we analyzed that did contain both LEDs required, such as the Apple Watch or various FitBits, the API access was limited to high level interpretations of biometric data from the user. Metrics like sleep quality, step counts, or heart rate were provided but access to the low level data was not. In order to adequately analyze the quality of PPG traces being received from a wrist-worn pulse oximeter, we we built our own wrist-worn device to measure SpO2. The sensors used and devices created for the purposes of experiments are described in the remainder of this section.

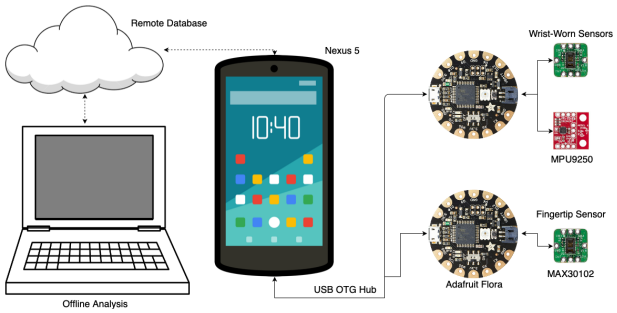


Figure 5: The data collection platform and sensors.

### 4.1.1 MAX30102 Sensor

We use the MAX30102 [10] reflective pulse oximeter from Maxim Integrated for our data collection. The sensor's provided by Maxim Integrated are the same as those commonly used by manufacturers such as Samsung for oxygen saturation measurement on smart phones and other devices. The sensor is described by the manufacturer as *an integrated pulse oximetry and heart-rate monitor biosensor module*. It provides red and infrared source LED's onboard the chip with an adjacent photodetector. Communication with microcontrollers is accomplished via the $I_2C$ protocol and where the sensors publish readings with a sample rate of 25Hz.

### 4.1.2 Wearable Prototype

In order to measure $SpO_2$ from a user with confidence, sensors are used to take measurements from two points of contact on a single user, namely the fingertip and the wrist. During experiments, PPG traces and other data are collected from all sensors simultaneously.

#### 4.1.3 Wrist-Worn Reflective Sensor with Motion Tracking

This is the primary sensor platform. The wearable consists of two sensors, including the MAX30102 sensor described in section 4.1.1, and an MPU9250 IMU sensor to track acceleration and rotation of the wrist worn device. Readings from the two sensors are captured and aligned using an Adafruit FLORA microcontroller. The three components are sewn into a fitness band for stability and consistency across measurements. The wrist-worn device is attached to the dominant hand of a user during experiments. The device allows for users to maintain range of motions in their wrist and movement throughout the duration of experiments is encouraged. The implications of using the methods described in this paper on a custom device versus a consumer grade device are discussed in section 6. 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.


Figure 6: Custom Wrist Wearable and Sensor Bed

#### 4.1.4 Fingertip-Worn Reflective Sensor

To establish a baseline for best-case signal from the MAX30102 sensor, we attach a second sensor to the index fingertip of the non-dominant hand of the user. The sensor is attached with medical tape to ensure a consistently applied pressure. The signal is again captured using an Adafruit FLORA microcontroller.

Our ground truth uses the exact same sensor applied to both the wrist and fingertip. This eliminates a variable in the experiment: we are interested in reliability across different measurement sites (specifically, the wrist vs fingertips), rather than across different hardware manufacturers. The MAX30102 was demonstrated to be a reliable sensor for measuring signals from the fingertip in section 2.

#### 4.1.5 Collecting and Aligning Sensor Signals

We wrote a custom Android application to capture and visualize signals from all sensors. The Android application communicates with each device over the USB serial protocol. A USB hub is used to communicate with the devices simultaneously as well as to provide power to each device. To later align the readings between devices, a timestamp is attached by the Android application when each reading is received. Finally, the application saves the collected readings to a remote database for offline processing.

### 4.2 Software

In addition to the Android application used to visualize data streams, we developed several Python applications to clean, align, and transform incoming data to be used with various out of the box machine learning libraries.

#### 4.2.1 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 use two radiance signals from the red and infrared LEDs, and two signals from the

magnitude of the gyroscope and accelerometer in the MPU9250. As discussed in section 7, it has been shown that motion of the device can be used to detect noise in the PPG reading. The magnitude of various motion readings from the IMU are used to automatically filter motion artifacts during classification.

We use the Tsfresh [9] Python library to extract features from our time series data. The library automatically selects features by calculating a comprehensive set of features on the provided data and then pruning the list of features based on the provided labels being predicted. Feature significance is performed using the Benjamini Hochberg procedure [11]. Depending on the training data provided, approximately 900-1000 features are selected by the library. Many features have a very low significance and can be pruned without affecting classifier performance. Table 1 shows the top 15 features of the 1000 total features. The first argument is the channel used, and additional parameters are dependent on the feature extracted. Detailed descriptions of the features and the Python API used to generate them are available at [4]. Based on these tests, the infrared LED channel adds the most information with a smaller dependency on the red LED and gyroscopic magnitude channels.

Table 1: Top 15 features.

| **Tsfresh Function Call** |
|---|
| longest_strike_below_mean('ir') |
| autocorrelation('ir', 6) |
| autocorrelation('ir', 5) |
| autocorrelation('ir', 7) |
| autocorrelation('ir', 8) |
| autocorrelation('ir', 9) |
| cid_ce('ir', normalize=True) |
| autocorrelation('ir', 4) |
| ar_coefficient('red', {"coeff": 0, "k": 10}) |
| spkt_welch_density('red', "coeff": 2) |
| ar_coefficient('ir', {"coeff": 0, "k": 10}) |
| mean('gyro') |
| sum_values('gyro') |
| fft_coefficient('gyro', {"coeff": 0, "attr": "abs"}) |
| fft_coefficient('gyro', {"coeff": 0, "attr": "real"}) |

Table 2: Optimal XGBoost parameters.

| **Parameter** | **Value** |
|---|---|
| Learning Rate | 0.1 |
| Estimators | 100 |
| Max Depth | 3 |
| Minimum Child Weight | 3 |
| Regularization Alpha | 0.3 |
| Subsample Ratio | 0.9 |
| Objective | Logistic Binary |

To select the classifier that best generalizes to unseen data, we compare several binary classifiers available in the scikit-learn library [12] using multiple validation sets. We found gradient boosting classifiers to provide robustness, generalizability, and the most consistently usable results. We use the XGBoost library [13], as it provides similar results with faster training times. We perform a cross validated grid search of hyperparameters to further tune the classifier. Evaluating performance based on data trained across several individuals yields the optimized parameters shown in Table 2. Data was trained using 10-fold cross validation, individual folds were checked against multiple separate validation sets.

During training in all experiments, non-overlapping windows are used to ensure that feature data is independent. Not only does preventing overlap ensure data remains i.i.d., but it also reduces feature extraction and training time by a factor equal to the windowing size, which is 100 in a majority of experiments. When applying the trained classifiers to unseen data, sliding windows of features with a step-size of 1 are taken. This ensures that we have the maximum number of output results with which to analyze.

### 4.3 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. Table 3 shows a summary of the 10 participants. Users range from 20-55 years of age and vary greatly in skin colour. 18000 readings are used from each user, corresponding to 12 minutes of readings acquired at a rate of 25Hz.

[4] https://tsfresh.readthedocs.io/en/latest/text/list_of_features.html

Table 3: Participant Information and proportion of reliable labels for each.

| User | Age | Relative Skin Tone | Proportion of reliable readings (within $\pm 2$ of fingertip) |
|---|---|---|---|
| 1 | 28 | Light | 47.3% |
| 2 | 24 | Light | 17.6% |
| 3 | 32 | Dark | 72.0% |
| 4 | 38 | Light | 44.8% |
| 5 | 20 | Dark | 0.3% |
| 6 | 31 | Medium | 28.3% |
| 7 | 24 | Dark | 2.1% |
| 8 | 31 | Medium | 6.0% |
| 9 | 26 | Light | 7.3% |
| 10 | 55 | Light | 16.5% |

We also how how much of the data collected from the wrist sensor is within 2 percentage points of the fingertip sensor. This shows that without adequate filtering, generally, very few readings from the wrist are accurate. Also notable, is the drastic variation between users in the proportion of readings considered reliable. These differences could stem from a large variety of variables that cannot be controlled in the wild. Variables such as; skin colour, device tightness, wrist thickness, movement, and ambient light, can all affect how much of a signal collected from a user is reliable.

## 5 Experimental Evaluation

We evaluate the ability of WristO$_2$ to correctly identify clean PPG readings for the purposes of measuring peripheral oxygen saturation. Through our data analysis we answer the following questions:

1. How does WristO$_2$ perform compared with existing algorithms?
2. How effective is WristO$_2$ across different measurement sites?
3. Does classification translate to unseen skin tone colours?
4. Can we calibrate WristO$_2$ on a per-user basis?
5. How does the IMU effect performance?
6. How do domain-specific hyper-parameters effect performance?

Our main performance metric is the root mean square error (RMSE) of readings taken from the wrist, compared to the fingertip ground truth. We also evaluate the precision of the WristO$_2$ classifier, as defined in Section 3.

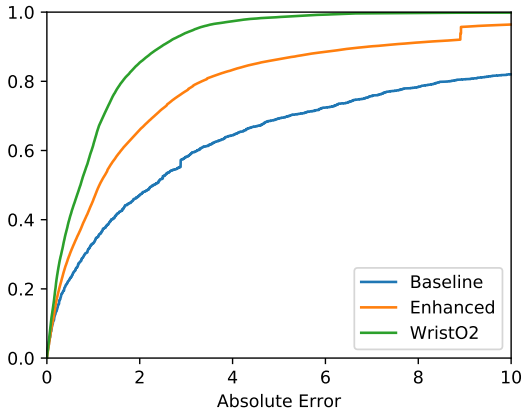


Figure 7: CDF of RMSE for existing algorithms and WristO$_2$ for all users.

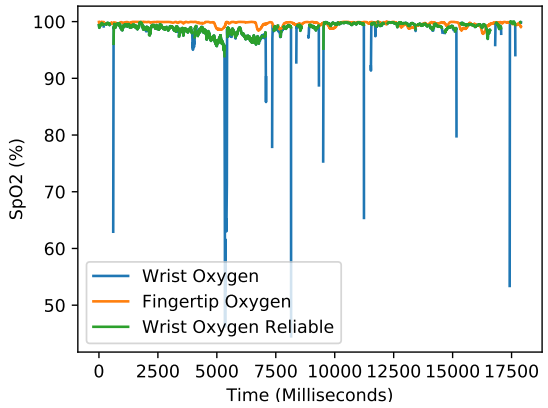


Figure 8: Trace of existing algorithms and WristO$_2$ for a single user.

Table 4: Mean (Std. Dev.) of classification results across all 10 users, trained and tested on the top of the wrist.

| **WristO$_2$ Precision** | **Baseline RMSE** | **Enhanced RMSE** | **WristO$_2$ RMSE** |
|---|---|---|---|
| 73% (19%) | 14.5% (6.9%) | 6.7% (4.4%) | 1.5% (0.7%) |

### 5.1 Performance of WristO$_2$

We perform leave-one-out cross validation across participants, where a single participant's signal is classified given training data from all others. Figure 7 shows the resulting absolute errors of wrist sensor readings across all readings for both existing algorithms and WristO$_2$, and Table 4 shows a summary of classification and algorithm performance. Pruning results with WristO$_2$ shows a drastic reduction in error compared to existing methods. WristO$_2$ reduces RMSE of $SpO_2$ measurement by an order of magnitude compared to the baseline algorithm, and by more than 4 times for the enhanced algorithm.

Figure 8 shows a trace for a single user, to better illustrate the effect of WristO$_2$, The blue line representing the enhanced algorithm applied to a PPG trace collected from the wrist over 12 minutes, and the orange line representing the enhanced algorithm applied to the signal collected from the fingertip during the same session. Finally, the green line in Figure 8 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.

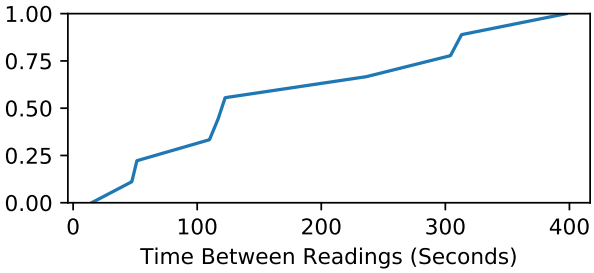


Figure 9: CDF of longest delay between valid readings.

The reduction in error comes at the cost of producing less readings compared to the existing algorithms. Figure 9 shows the CDF of the maximum size of an interval with no reliable values for all users across the leave-one-out validation. In other words, the longest silent window for each user where the classifier is returning no reliable readings. The worst case scenario for a no reading window was approximately 6 minutes and 40 seconds, while the average worst-case across all users is approximately 3 minutes. Given that the method for acquiring reliable readings currently 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.

### 5.2 Comparing Measurement Sites

Each participant listed in table 3 participated in a second trial where the wrist-worn pulse oximeter was applied to the bottom of the wrist, rather than the top. Using the same approach described in section 5.1, we run leave-one-user-out cross-validation to test the performance of WristO$_2$ with a signal collected from the bottom of the wrist. Table 5 presents the results.

Table 5: Mean of classification results across all 10 users for varied measurement sites.

| **Training Site** | **Testing Site** | **WristO$_2$ Precision** | **Enhanced RMSE** | **WristO$_2$ RMSE** | **Participants with no readings** |
|---|---|---|---|---|---|
| Bottom of Wrist | Bottom of Wrist | 23% (33%) | 14.7% (13.1%) | 12.0% (3.7%) | 5 |
| Top of Wrist | Bottom of Wrist | 25% (28%) | 14.7% (13.1%) | 8.9% (10.3%) | 2 |

The 10 users had an average RMSE of 14.68% when measurements were taken from the bottom of the wrist with the enhanced algorithm. After pruning using the classifier trained on this data, the average RMSE is reduced to 12.0%.

Despite the slight improvement, it should be noted that for half of the users, WristO$_2$ pruned all values, meaning no readings in the 12 minute window were marked as reliable.

Applying WristO$_2$ to the bottom of the wrist when trained on signals from the top shows improved. The original 14.68% RMSE is further reduced to 8.84%. Furthermore, only two out of the ten trials in this instance provided no reliable readings. So, given a situation where data must be collected from the bottom of the wrist, the original classifier trained in section 5.1 can still be used to prune less reliable labels. It is also notably less aggressive in pruning than the classifier trained on readings from the bottom of the wrist.

It is clear that it better to collect traces, and train classifiers using data acquired from the top of the users wrist. This is a positive result when considering that this is where almost all consumer grade devices choose to collect data from already.

### 5.3 Effect of Skin Tone

As discussed in section 7, 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. This section aims to quantify potential difficulty in collecting reliable PPG traces from users of various skin tones. Five of the participants had skin colour that we qualitatively define as light, relative to the other users. We separate the users qualitatively into two groups, lighter- or darker-skinned, and train classifiers with all permutations of these 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 6: Effects of skin tone with various training and testing permutations.

| **Training Group** | **Testing Group** | **WristO$_2$ Precision** | **Enhanced RMSE** | **WristO$_2$ RMSE** |
|---|---|---|---|---|
| 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%) |

Table 6 shows that precision is improved when classifying on lighter skin as opposed to darker skin, regardless of the skin-tone used during training. Unexpectedly, results are slightly improved for predicting lighter skin signal reliability when the darker skinned group was used for training. Given the high variance in results, it is likely that this discrepancy is due to the small sample size. There is also slightly less data with the light-to-light experiment since leave-one-out cross validation is used. Utilizing 4 users for training, instead of 5 for the dark-light experiment, means less data is available for training and performance could be affected.

We caution that sample size is too small to draw strong conclusions about the magnitude of effects, and much more data will be needed to prove 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.

### 5.4 Per-User Training

This experiment attempts to show the viability of WristO$_2$ to provide on-the-fly training to build a personalized 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 smart watch, and a similar fingertip sensor such as those that exist in the back of various Samsung smart phones. 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, aligned, and preprocessed, the classifier can be retrained with the additional data to provide the user with more reliable readings from the wrist-worn device.

We train a classifier with 9 users and test it on an unseen user. We then retrain the classifier with 2 minutes of additional calibration data, and again with 10 minutes. The results are summarized in Table 7. 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 specific training data on top of the original training set further reduces the RMSE up to 0.7%.

Table 7: Adding user calibration data to increase WristO$_2$ performance.

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

### 5.5 Importance of Accelerometer and Gyroscope

To study the effect of the features extracted from the IMU signal on classification, we run a similar experiment to section 5.1 with varied combinations of features from the LEDs and IMU sensor. Table 8 summarizes the results of the experiments

Table 8: Effects of IMU features on classification.

| Signal Channels | Num. Features | WristO$_2$ Precision | Enhanced RMSE | WristO$_2$ RMSE |
|---|---|---|---|---|
| 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%) |

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 a very low 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.

### 5.6 Effects of varied thresholds and window sizes

This section aims to tune two parameters discussed in section 3, namely the window size for feature extraction, and the threshold of reliability.

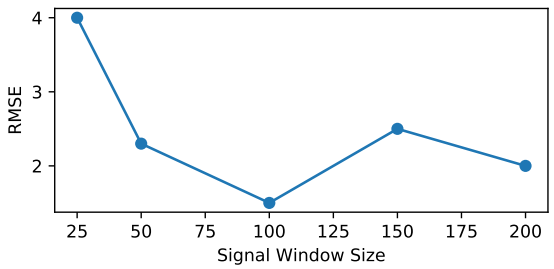


Figure 10: RMSE for different signal window size.

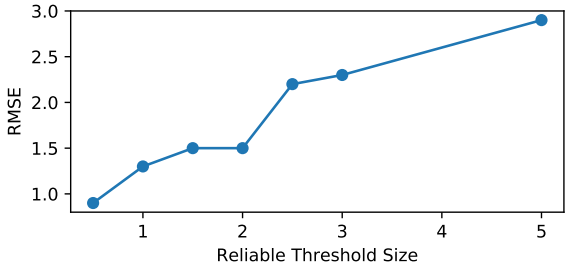


Figure 11: RMSE for different thresholds.

#### 5.6.1 Window Size

Figure 10 shows varied window sizes and their resulting RMSE with a window size of 100 providing the lowest. However, it is useful to know that acceptable results can be achieved with smaller window sizes. This is useful if performance is an issue. Because the feature extraction takes longer with larger window sizes, we can use smaller feature windows at less frequent intervals to decrease feature extraction time with only a small performance penalty.

#### 5.6.2 Reliability Threshold

Figure 11 shows varied sizes of reliability thresholds 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 12 demonstrates for the varied threshold sizes.

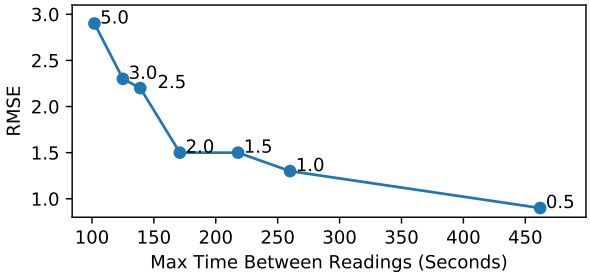


Figure 12: The trade-off between quantity and quality of readings for different reliability thresholds. For every threshold, the X axis shows the resulting worst-case interval between reliable readings, and the Y axis shows the resulting RMSE.

## 6 Discussion

This section covers future work for Wrist$O_2$, including necessary steps for deploying it in the wild on consumer grade devices.

**Reducing Compute Costs** For Wrist$O_2$ to be deployed on existing wrist-worn mobile devices, measures must be taken to ensure the performance and battery life of mobile devices are not affected. First, the feature set could be pruned far enough that the computational cost of feature extraction is reasonable to perform on live data directly onboard the device taking the measurements. In addition to pruning features, work similar to Sidewinder[14] could be used to offload signal reliability calculation to a lower powered processor, and subsequently wake the device when a usable signal is detected.

Alternatively, instead of optimizing the feature extraction, we could utilize a cloud service to stream the collected data, offload the computation, and collect the results. The data is small enough that an hours worth of data can be transferred within seconds to a remote server that processes the data.

**Improving Classification** Although the classifier is trained, tested, and validated on a diverse group of people, the small number of people used in the study could be limiting the ability of the classifier to predict some values if they are all within a healthy range. Future work will include more participants, and incorporate participants with lower $SpO_2$.

We are also considering extending the classifier to multi-label classification or regression. That is, predict not whether a signal will produce a reliable label within a certain threshold, but predict the confidence that the label will be produced within various different thresholds. For example 1%, 3%, and 10%. We leave this to future work.

**Extending to Existing Wearable Devices** The next obvious iteration of the wrist-worn device is to either build or utilize a consumer grade device. Wrist$O_2$ could be applied in consumer grade devices if low level access to the LED sensor were to be provided. The use of a consumer grade device could potentially improve results solely based on the quality of the hardware.

## 7 Related Work

To our knowledge this is the first work that applies state-of-the-art feature extraction and machine learning approaches to increase the reliability of $SpO_2$ measurements taken from the signals of wrist-worn devices.

Much of existing work on reflective sensors focused on heart rate measurement, such as rule based detection of heart rate for reliability [15]. Ra et al. perform reliability detection in the context of wrist heart-rate measurement using hidden Markov models applied to a single LED source on existing smart watches [16]. There has been work done to improve reading reliability in fingertip sensors at the algorithmic level for both heart rate and $SpO_2$ through signal preprocessing and noise reduction [6][17]. Possible wearability sites, including the wrist, and various sensor configurations have been considered in the context of telehealth monitoring [4, 18]. Other work has documented the process of building transitive pulse oximeters from scratch [19]. Reflective pulse oximeters are widely used and studied in medicine in places where transitive pulse oximeters are not feasible, such as infant monitoring [20].

Accuracy and reliability of fingertip worn pulse oximeters have been analyzed in great detail, such as quantifying quality of $SpO_2$ measurements in patients with specific conditions or qualities. Severinghaus et al.[21] showed that bias in $SpO_2$ measurements increases during a state of anemia (low red blood cell count). Emery et al. [22] and Cote et al. [23] showed the effects of dark skin pigmentation and ink in convoluting measurements of fingertip worn pulse

oximeters. Additionally Lee et al. [24] showed that lower true pulse oximetery values were overestimated for a specific set of people from Singapore due to darker pigmentation.

Yao et al.[25] used simple motion sensing to remove noise from movement artifacts to improve signal reliability in ambulatory environments. Yan et al.[26] used a more sophisticated feature extraction to remove motion and other noise artifacts in the context of at home fingertip pulse oximeters used for telehealth monitoring.

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 [27]. Although they currently do not employ $SpO_2$ in their consideration of patient health, this project could aid their work by providing a reliability measure for $SpO_2$ readings.

## 8 Conclusion

In this work we study the reliability of $SpO_2$ measurements from a wrist-worn pulse oximeter, and show that existing algorithms do not provide reliable readings. We propose Wrist$O_2$, 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 Wrist$O_2$, we show that we can reduce error in the measurements taken from the wrist by up to an order of magnitude. Additionally we demonstrate that even after pruning results, the frequency of reliable readings is still high enough to be useful, and in fact significantly better than current methods that require user intervention. We discuss the effects of sensor placement and skin tone on Wrist$O_2$, explore the effect of IMU information, and propose platforms for user level calibration. Finally, we discuss next steps to deploy this technique in the wild. We present this research as a proof of concept for manufacturers and developers to implement reliable data collection platforms with which to build useful applications based on peripheral oxygen saturation.

## References


[1] Robert T Brouillette, Angela Morielli, Andra Leimanis, Karen A Waters, Rina Luciano, and Francine M Ducharme. Nocturnal pulse oximetry as an abbreviated testing modality for pediatric obstructive sleep apnea. *Pediatrics*, 105(2):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. *European Respiratory Journal*, 7(2):274–278, 1994.

[3] Andreas H Taenzer, Joshua B Pyke, Susan P McGrath, and George T Blike. Impact of pulse oximetry surveillance on rescue events and intensive care unit transfersa before-and-after concurrence study. *Anesthesiology: The Journal of the American Society of Anesthesiologists*, 112(2):282–287, 2010.

[4] Y Mendelson and C Pujary. Measurement site and photodetector size considerations in optimizing power consumption of a wearable reflectance pulse oximeter. In *Engineering in Medicine and Biology Society, 2003. Proceedings of the 25th Annual International Conference of the IEEE*, volume 4, pages 3016–3019. IEEE, 2003.

[5] Yitzhak Mendelson and Burt D Ochs. Noninvasive pulse oximetry utilizing skin reflectance photoplethysmography. *IEEE Transactions on Biomedical Engineering*, 35(10):798–805, 1988.

[6] P Madhan Mohan, A Annie Nisha, V Nagarajan, and E Smiley Jeya Jothi. Measurement of arterial oxygen saturation (spo 2) using ppg optical sensor. In *Communication and Signal Processing (ICCSP), 2016 International Conference on*, pages 1136–1140. IEEE, 2016.

[7] Toshiyo Tamura, Yuka Maeda, Masaki Sekine, and Masaki Yoshida. Wearable photoplethysmographic sensors—past and present. *Electronics*, 3(2):282–302, 2014.

[8] Wei Chen, Idowu Ayoola, Sidarto Bambang Oetomo, and Loe Feijs. Non-invasive blood oxygen saturation monitoring for neonates using reflectance pulse oximeter. In *Proceedings of the Conference on Design, Automation and Test in Europe*, DATE '10, pages 1530–1535, 3001 Leuven, Belgium, Belgium, 2010. European Design and Automation Association.

[9] Maximilian Christ, Nils Braun, Julius Neuffer, and Andreas W Kempa-Liehr. Time series feature extraction on basis of scalable hypothesis tests (tsfresh–a python package). *Neurocomputing*, 2018.

[10] Maxim Integrated. MAX30102 high-sensitivity pulse oximeter and heart-rate sensor for wearable health, 2018.

[11] Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188, 2001.

[12] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[13] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM.

[14] Daniyal Liaqat, Silviu Jingoi, Eyal de Lara, Ashvin Goel, Wilson To, Kevin Lee, Italo De Moraes Garcia, and Manuel Saldana. Sidewinder: An energy efficient and developer friendly heterogeneous architecture for continuous mobile sensing. *ACM SIGARCH Computer Architecture News*, 44(2):205–215, 2016.

[15] Ammar Al Ali, Divya S Breed, Jerome J Novak, and Massi E Kiani. Pulse oximetry data confidence indicator, January 27 2004. US Patent 6,684,090.

[16] Ho-Kyeong Ra, Jungmo Ahn, Hee Jung Yoon, Dukyong Yoon, Sang Hyuk Son, and JeongGil Ko. I am a "smart" watch, smart enough to know the accuracy of my own heart rate sensor. In *Proceedings of the 18th International Workshop on Mobile Computing Systems and Applications*, HotMobile '17, pages 49–54, New York, NY, USA, 2017. ACM.

[17] Jianchu Yao and Steve Warren. A short study to assess the potential of independent component analysis for motion artifact separation in wearable pulse oximeter signals. In *Engineering in Medicine and Biology Society, 2005. IEEE-EMBS 2005. 27th Annual International Conference of the*, pages 3585–3588. IEEE, 2005.

[18] Y Mendelson, RJ Duckworth, and G Comtois. A wearable reflectance pulse oximeter for remote physiological monitoring. In *Engineering in Medicine and Biology Society, 2006. EMBS'06. 28th Annual International Conference of the IEEE*, pages 912–915. IEEE, 2006.

[19] Sangeeta Bagha and Laxmi Shaw. A real time analysis of ppg signal for measurement of spo2 and pulse rate. *International journal of computer applications*, 36(11):45–50, 2011.

[20] David R Tobler, Mohamed K Diab, and Robert J Kopotic. Fetal pulse oximetry sensor, September 4 2001. US Patent 6,285,896.

[21] John W Severinghaus and Shin O Koh. Effect of anemia on pulse oximeter accuracy at low saturation. *Journal of clinical monitoring*, 6(2):85–88, 1990.

[22] JR Emery. Skin pigmentation as an influence on the accuracy of pulse oximetry. *Journal of perinatology: official journal of the California Perinatal Association*, 7(4):329–330, 1987.

[23] Charles J Coté, E Andrew Goldstein, William H Fuchsman, and David C Hoaglin. The effect of nail polish on pulse oximetry. *Anesthesia and analgesia*, 67(7):683–686, 1988.

[24] KH Lee, KP Hui, WC Tan, and TK Lim. Factors influencing pulse oximetry as compared to functional arterial saturation in multi-ethnic singapore. *Singapore medical journal*, 34:385–385, 1993.

[25] Jianchu Yao and Steve Warren. A novel algorithm to separate motion artifacts from photoplethysmographic signals obtained with a reflectance pulse oximeter. In *Engineering in Medicine and Biology Society, 2004. IEMBS'04. 26th Annual International Conference of the IEEE*, volume 1, pages 2153–2156. IEEE, 2004.

[26] Yong-Sheng Yan and Yuan-Ting Zhang. An efficient motion-resistant method for wearable pulse oximeter. *IEEE Transactions on information technology in biomedicine*, 12(3):399–405, 2008.

[27] Daniyal Liaqat, Ishan Thukral, Parco Sin, Hisham Alshaer, Frank Rudzicz, Eyal de Lara, Robert Wu, and Andrea Gershon. Poster: Wearcopd - monitoring copd patients remotely using smartwatches. In *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services Companion*, MobiSys '16 Companion, pages 139–139, New York, NY, USA, 2016. ACM.