Reliable Peripheral Oxygen Saturation Readings from Wrist-Worn Pulse Oximeters

Caleb Phillips
caleb@cs.toronto.edu

Daniyal Liaqat
dliaqat@cs.toronto.edu

Moshe Gabel
mgabel@cs.toronto.edu

Eyal De Lara
delara@cs.toronto.edu

***Abstract*— Off the shelf wrist-worn devices, such as the Apple Watch, FitBit, and Samsung Gear, have an onboard sensor called a pulse oximeter that is generally used to measure heart rate. Oxygen saturation readings are a vital measurement in health care, fitness, and many other domains. Although pulse oximeters have the ability to measure oxygen saturation, they are rarely employed to do so in wrist-worn devices. Our experiments show that collecting this biometric data from wrist-based devices can lead to over 90% of readings being inaccurate and unusable. After implementing a pipeline to collect and analyze signals from various participants, we develop a classifier that can prune a majority of unreliable signals and improve the overall quality of readings produced. After pruning values deemed unreliable by the classifier, initial results show anywhere from an eight to forty fold reduction in errors between the wrist mounted sensors and a ground truth device, including users for which the classifier was not trained. It is the hope that this work will allow manufacturers to use these devices to obtain highly reliable measures of oxygen saturation, and promote developers to built interesting and useful applications on top of more confidently acquired data.**

I. Introduction

Advances in continuous health monitoring could be used as a helpful aid for physicians treatment and diagnosis of patients. At home health monitoring promotes a way of mitigating exorbitant health care incurred costs by allowing at home diagnosis and management of diseases. Personal health tracking devices and software are used widely by average consumers, athletes, pilots, and astronauts alike. A pulse oximeter is a common sensor used in all of these domains to measure heart rate, and less commonly, peripheral oxygen saturation (SpO_2) of the blood.

Oxygen saturation readings have many use cases in healthcare monitoring if collected with confidence, and as such are a primary vital sign used by nurses and physicians to monitor patients. 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]. And for personal health tracking, cell phone manufacturers have recently provided an onboard pulse oximeter on the back of smartphones that can provide an instantaneous SpO_2 reading given contact with a users fingertip.

Unlike heart-rate biometrics, peripheral oxygen saturation does not change drastically over time in most healthy individuals, who range from 95% to 100% oxygen saturation without symptoms. Hypoxemia is a level of oxygen in the blood of 90% or less, for which medical attention is suggested. And for people with severe COPD, asthma, or other breathing related conditions, it can be useful to track potentially sudden changes in SpO_2 levels in order to provide treatment recommendations or contact emergency services if levels fall below this threshold. Readings for these individuals and average consumers alike are currently collected either continuously by attaching a transitive pulse oximeter to the users fingertip, or intermittently by holding a finger against the surface of a reflective pulse oximeter on a smartphone. These methods either inhibit the use of the wearers hand, are uncomfortable to have attached to a fingertip for long duration's (as attested to by some of our users), or require constant input and attention from the user.

In this paper, we aim to demonstrate that an intermittent reliable SpO_2 signal can be taken automatically from a users wrist using the same sensors currently employed in existing wrist-worn devices, such as the Apple Watch, FitBit, and Samsung Gear. We develop a wrist worn sensor collection platform in order to access the raw biometric signal required to calculate a peripheral oxygen saturation reading. Alongside the wrist-worn device we attach both an off the shelf reflective sensor and a transitive fingertip device to compare signals and establish a ground truth measurement between multiple reference points. Finally, using this experimental platform we develop a classifier to identify a signals ability to provide an accurate SpO_2 reading to the user with high precision.

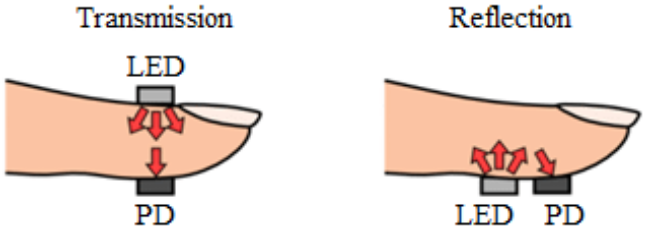
II. Background

Pulse oximeters come in two primary types, the first is *transitive* which is often used for medical monitoring purposes within hospitals and applications where highly accurate measurements are desired. Although these devices are highly accurate, they require a device to be clipped to the wearer (usually a finger), can impede the wearers use of their hands, and are generally not well suited for continuous monitoring where the user requires mobility.

The second type of pulse oximeter is *reflective*[4], the sensors for which can be embedded in everyday devices

such as a watch or ring. Although these sensors are more easily worn for long periods and can be integrated into other devices, their SpO_2 measurements are generally not as accurate and more frequently unreliable [5].

Fig. 1. Photodetector and LED Placement for Pulse Oximeters



Pulse oximeters in general are desirable tools for monitoring patient health because they do so in a non-invasive manner. Both transitive and reflective sensors operate by emitting red and infrared wavelengths of light from LED's, and measuring the interactions of the wearers tissue with these light sources via a photodetector. Their difference lies in the positioning of the photodetector relative to the LED's. A transitive pulse oximeter has the photodetector sit opposite the LED source, and measures the light interaction with tissue as it *transits* the tissue. Transitive devices include the highly accurate fingertip worn devices commonly used in hospital settings, but can be worn on any body part with lots of blood flow and thin enough tissue.

A reflective pulse oximeter has a photodetector that sits directly beside the LED light sources, and measures the interaction of the red and infrared LED's as they *reflect* off of the wearers tissue. These devices are common in watches and cell phones, and although they can potentially provide accurate results, they are not nearly as reliable as their transitive counterparts. Because they are not always directly in contact with and applying pressure to the skin via a fingertip clamp, they are more susceptible to noise from ambient light, motion, and pressure changes. A reflective pulse oximeter is also not limited to a specific measuring site, however locations with a high blood profusion, such as the wrist, earlobe, or forehead are commonly used. Although reflective pulse oximeters are widely used today, their capacity is 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 measure SpO_2.

A pulse oximeter requires a red and infrared LED to measure interactions of the light with red oxygen saturated blood and other body tissue. A pulse oximeter measures SpO_2 readings by producing a photoplethysmogram (PPG) of the incoming red and infrared light sources. An estimate of oxygen saturation is produced by calculating a ratio of ratios between the reflection or transmittance of the red and infrared colours in the PPG, as described by the following equation.

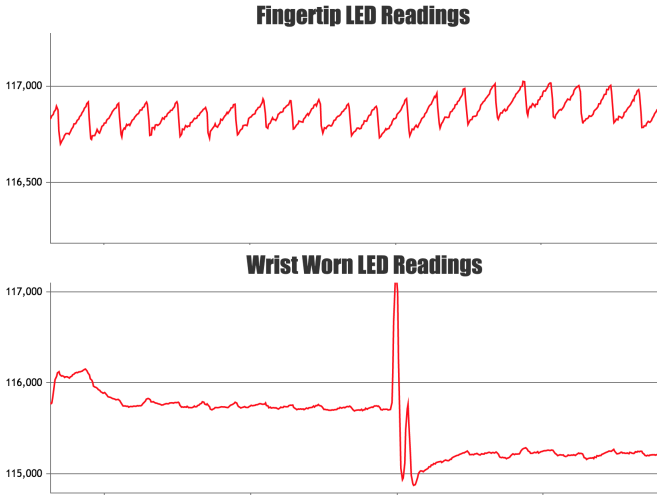
$$SpO_2 = y_0 - m \times \left(\frac{AC_{Red}/DC_{Red}}{AC_{IR}/DC_{IR}} \right)$$

Although the details of this equation are not important it is useful to note that the y_0 and m terms represent a linear fit for device readings, and would generally come from the manufacturer of the specific sensor after they have calibrated the sensor against a ground truth. The equation is based on Beer-Lamberts law, and described more thoroughly in [6].

A. PPG Traces

The noise incurred with measuring a PPG trace from a wrist worn device can be seen in figure 2. The first image shows a typical single channel PPG trace captured from a fingertip worn reflective pulse-oximeter over several seconds. The signal can be seen to capture the change in flow of oxygenated blood through the fingertip. The second image shows a trace from the same user, during the same time, using the same sensor, taken from the wrist of the user. The signal, even with a clean contact to the users skin, produces a much noisier signal. The spike in the middle demonstrates either a motion or ambient light artifact, where poor contact with the skin or a users movements can cause errors and discontinuity in the signal. Algorithms used to produce SpO_2 readings from a wrist-worn sensor must be able to mitigate and compensate for these errors.

Fig. 2. PPG Trace for a fingertip vs wrist attached sensor



III. Experimental Setup

This section describes the experimental setup and three wearable devices employed in it. The primary device is a wrist-worn sensor bed designed to collect signals that would resemble those collected from existing wrist-worn consumer electronics. It is the goal of this work to improve the quality of output produced by this device, while using the other sensors as reference and ground truth. All other infrastructure is designed to collect, align, and analyze the data collected from these three devices.

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. And in doing so a majority of 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 the level of access that manufacturers provide via their APIs to information such as the reflectance level of LEDs, which are needed to construct a 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 were forced to construct our own devices to allow for comparison and analysis. The sensors used and devices created for the purposes of experiments are described in the remainder of this section.

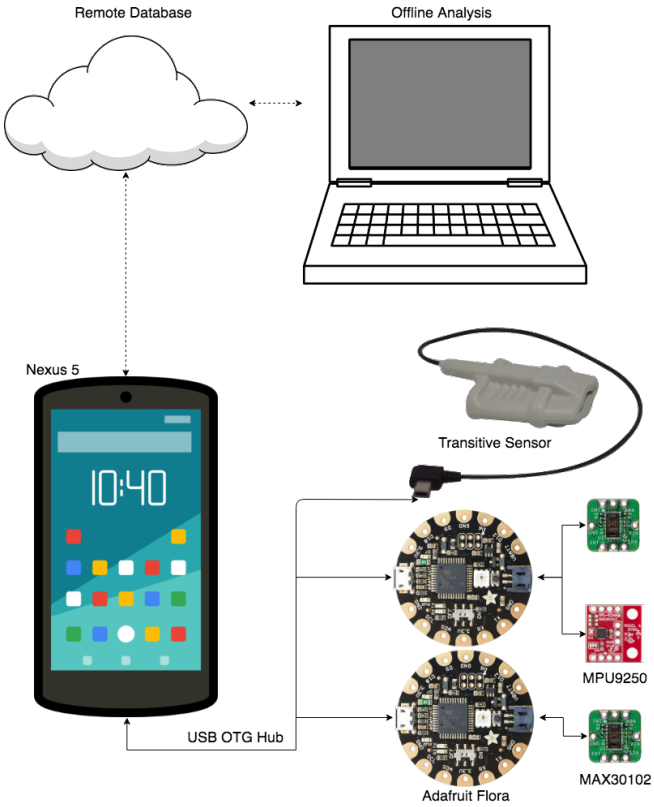


Fig. 3. Experimental Architecture

A. MAX30102 Sensor

We use the MAX30102[7] reflective pulse oximeter from Maxim Integrated for the purposes of creating custom wearable 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 publishes readings at a rate of 25Hz.

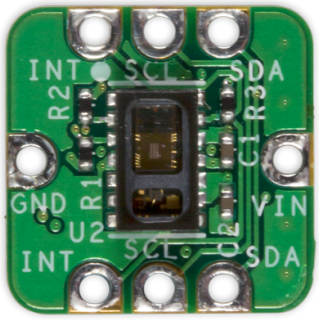


Fig. 4. MAX30102 Pulse Oximeter

B. Measurement Devices

In order to measure SpO_2 from a user with confidence, three separate devices are used to take measurements from three points of contact on a single user. Each device and its method of attachement to the user is described below. During experiments, PPG traces and other data are collected from all three simultaneously.

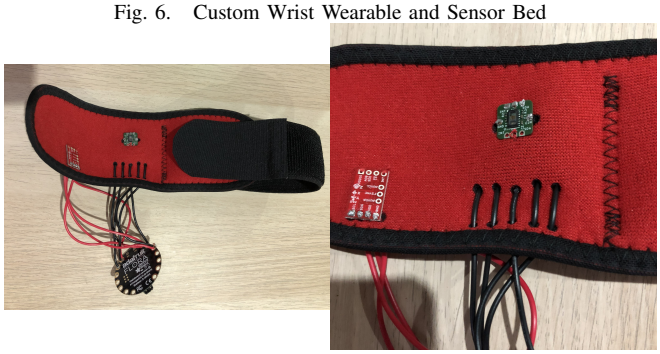
1) Wrist-Worn Reflective Sensor with Motion Tracking: This is the primary device within experiments for which we will show current shortcomings of the measurement site as well as provide an improved metric of SpO_2 reliability. The device consists of two sensors, including the MAX30102 sensor described in section III-A, 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 through the duration of experiments is encouraged. The fitness band used is larger than most consumer grade wrist-worn devices, and as such is possibly isolating more light from the sensor, and possibly removing ambient light artifacts that may otherwise be present. Although it is possible that this balances out against the MAX30102s lack of enclosure that would exist in a consumer grade device like an Apple Watch, as shown in figure 5. The implications of using the methods described in this paper on a custom device versus a consumer grade device are discussed in section VIII.

2) 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.

3) Fingertip-Worn Transitive Sensor: As discussed in section II, transitive sensors often provide higher levels of consistency and reliability in measuring SpO_2. To provide

Fig. 5. Enclosed Pulse Oximeters on underside of Apple Watch

Fig. 6. Custom Wrist Wearable and Sensor Bed



a third point of reference for measurement, a USB driven transitive pulse oximeter[8] is attached to the middle-finger of the non-dominant hand of the user. The USB device directly provides calculations of SpO_2. Given the reliability and consistency of transitive devices over reflective sensors, this device is desirable to confirm quality of incoming signals from the reflective pulse oximeters.

Fig. 7. Transitive Pulse Oximeter

C. Aligning Device Signals

To capture the signals from all three devices, a custom Android application is used. The Android application communicates to each device via 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.

D. Analyzing Data

In addition to the Android application used to visualize data streams, we develop several python applications to clean, align, and transform incoming data to be used with various out of the box machine learning libraries. In addition we have developed a web platform to visualize raw traces collected across experiments.

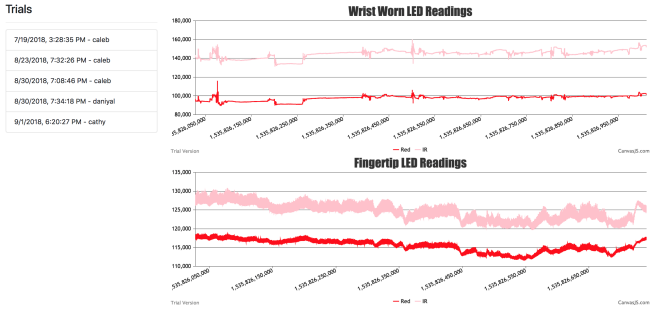


Fig. 8. PPG Web Platform

IV. Existing Algorithms

This section aims to quantify the current state of available algorithms for measuring SpO_2 from a reflective pulse oximeter.

Currently two implementations of the SpO_2 algorithm exist for the MAX30102 sensor we have employed. The first is the naive implementation supplied by the manufacturer for testing purposes of the device. The algorithm, although correct, provides a very primitive measure of reliability for SpO_2 calculations, and only relies on values being within a certain range to determine correctness.

The second algorithm is an enhancement of the first algorithm that implements the same calculation of SpO_2, however it relies on a less naive calculation of the incoming signals reliability. The code and a description of its algorithm is available through [9]. Notably, after performing baseline levelling of each signal, a Pearson correlation is calculated between the incoming red and infrared channels. This is a sensible approach since, despite the signals being within different ranges, their measurement site is the same and should therefore result in a highly correlated signal. To filter unreliable signals, this method calculates the Pearson correlation coefficient and discards any signals that produce a value below a calculated value of 0.4.

The code for both of the aforementioned algorithms are originally implemented for an Arduino capable board. For each Arduino program, we remove the Arduino relevant code, compile the remaining C code on a standard laptop, and apply the code offline to the traces collected from the devices containing the MAX30102 sensor, to allow for direct comparison between the two algorithms and general offline analysis.

A. Comparing Existing Algorithms to a Ground Truth

We analyze the ability of existing algorithms to produce accurate output when applied to signals retrieved from a wrist-worn device. To accomplish this, we compare the output of each algorithm applied to a signal retrieved from the wrist-worn device and a clean signal retrieved from the fingertip sensor.

A user in each experiment is seated. The non-dominant hand, to which the fingertip device is attached, is stationary for the duration. The dominant hand bearing the wrist-worn device is used moderately and performs normal activities such as typing and eating for the duration of the experiment. To provide a ground truth for true reliability of the SpO_2 calculation, we compare the algorithmic output for the wrist-worn device to the calculated output from the fingertip worn device. If the output from a given algorithm is within a threshold of $\pm 1.0\%$ from the fingertip value, we label the reading as reliable.

1) Maxim Algorithm: Figure 9 shows collected readings from the two devices for two different users. The red and blue lines in each user represent the SpO_2 output of the Maxim algorithm applied to the wrist-worn and fingertip sensors respectively. The readings are summarized in the table below with the first row providing information on the length of the trial. The second row indicates how many of the readings over the duration of the experiment were deemed reliable by the naive algorithm. And the final row indicates what percentage of the total number of readings for the wrist-worn device were with a threshold of $\pm 1.0\%$ from the fingertip device, indicating some level of confidence. It is obvious that across various users the naive implementation is far too lenient in its interpretation of what is a useable signal for calculating SpO_2.

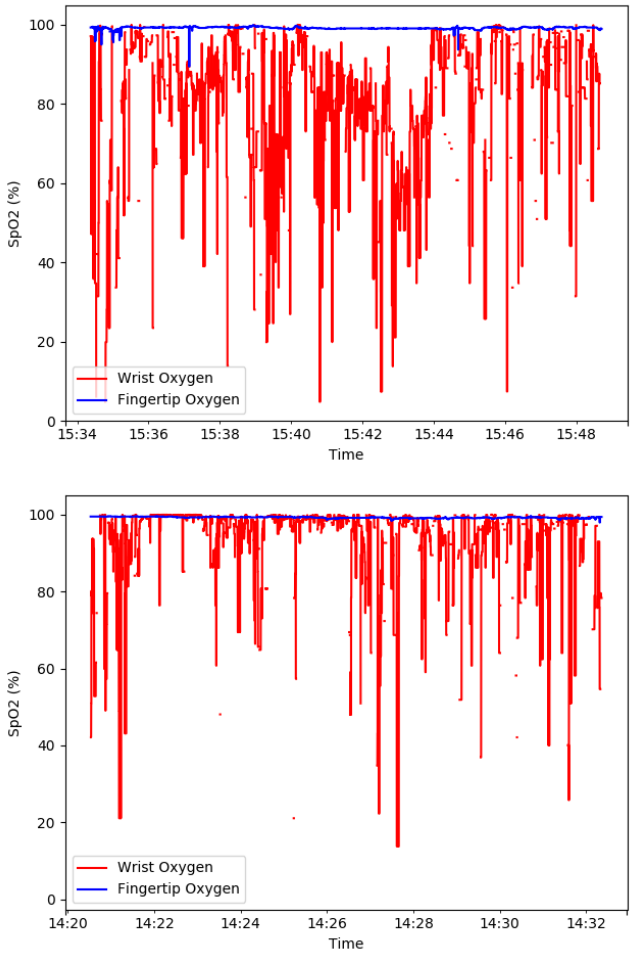
Maxim Algorithm	**User 1**	**User 2**
Duration (# Readings)	14m:24s (21602)	11m:53s (17837)
Valid by Algorithm	84.8%	78.5%
Within 1% of Fingertip	38.1%	5.6%

2) Enhanced Reliability Algorithm: Figure 10 shows the sample data across the same two users with the less naive *Enhanced* algorithm applied to the same data collected from the wrist-worn device. Although the reliability measure that the enhanced algorithm employs removes a majority of the unreliable readings from the output, anywhere from 59% to 97% of algorithmic readings are still greater than $\pm 1.0\%$ away from output collected from the fingertip trace. This of course varies depending on the user.

Enhanced Algorithm	**User 1**	**User 2**
Duration (# Readings)	14m:24s (21602)	11m:53s (17837)
Valid by Algorithm	6.4%	36.4%
Within 1% of Fingertip	0.2%	15%

The drastic differences between the experimental results across users 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 and useable.

Fig. 9. Maxim Algorithm on Wrist vs. Transitive Sensor, Two Users



B. Device Bias

Figure 11 shows traces of the fingertip reflective sensor vs the transitive sensor. It is clear from the plots that the readings from the 2 sensors have a bias of 2-3% between them, depending on the wearer. It is also obvious that the output of the reflective sensor shows results to a finer granularity. Both sensors seem to have errors. One or a combination of both of the devices needs to be used as ground truth in determining a level of reliability for the wrist-worn device. We analyze each possible label combination in section V-B.

V. Method

This section describes our method used to tighten the criteria for what is labelled as a reliable signal. The techniques are based on automated feature extraction and reliability classification using agreement between device labels. Although we still employ the original algorithms for calculating SpO_2, the goal is to only apply this algorithm to signal windows that we are sure will produce reliable readings.

Fig. 10. Enhanced Algorithm on Wrist vs. Transitive Sensor, Two Users

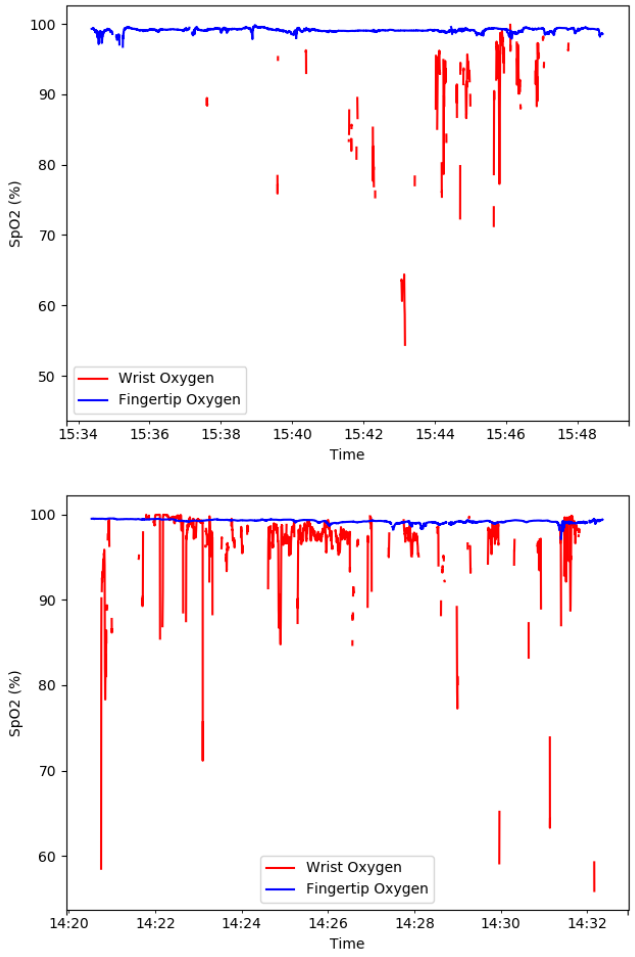
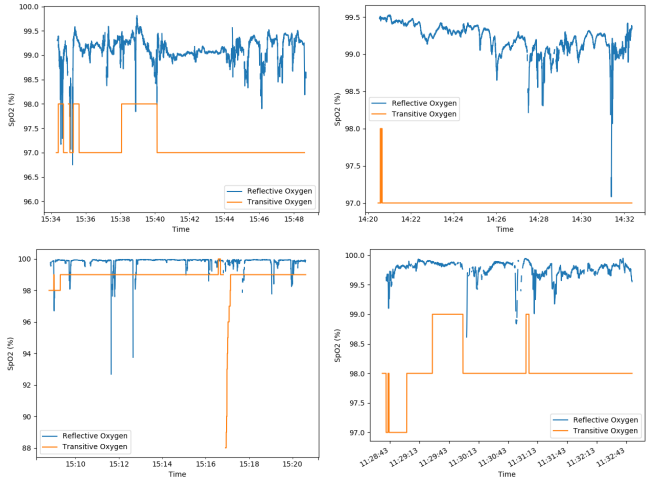


Fig. 11. Bias in Fingertip Reflective vs Transitive Sensor



A. Feature Engineering

Features are extracted from 4 continuous signals of the wrist worn device to use as inputs when predicting the reliability of the signal. Two signals from the radiance measured from the red and infrared LEDs, and two signals from the magnitude of the gyroscope and accelerometer in the MPU9250. As discussed in section IX, it has been shown that motion of the device can be used to indicate noise in the PPG reading. The magnitude of various motion readings from the IMU are used to automatically filter motion artifacts during classification.

The device readings are generally windowed to 4 seconds, or 100 readings at 25Hz, which corresponds to the same window size used by the algorithm discussed in section IV-A.2. We explore other potential window sizes in section VI-A.1. To extract features from the time series data, a brute force approach is taken with the python library tsfresh[10]. The library attempts to auto select features by calculating a comprehensive set of features on the provided data, and finally pruning the list of features based on the labels being predicted. The pruning removes features which provide no information based on the labels provided. Depending on the training data provided, approximately 900 features are selected by the library. After we train a classifier, further features are removed based on a uni-variate feature importance of 0 in the classifier chose. The final number of features calculated is approximately 450. Being that we only remove features that have a significance of 0, many features are left with a minuscule significance that can likely be pruned without affecting classifier performance. This step will also need to be taken to make features calculable on the fly in mobile devices without greatly impacting performance and energy consumption. We discuss potential solutions to this in section VIII. Furthermore, a majority of the motion features pruned are related to the acceleration, indicating that the gyroscope magnitude corresponds with motion artifacts more directly than acceleration. We eliminate the acceleration channel from all but these initial experiments without affecting performance.

~~B. Reliability Label~~

~~In order to develop a useful reliability classifier we need a confident reliability metric. We define a reliable signal as producing an output within some small threshold of one or more of the ground truth devices. To decide which set of devices to use as ground truth we compare several primitive classifiers. We test these classifiers with a threshold of 2.0, and further tune this parameter in section VI-A. When both fingertip devices are used as ground truth, the threshold is defined as the greatest distance between any two device readings, otherwise for a single fingertip device the threshold is just the distance between two valid readings. If any reading is deemed unreliable by the underlying SpO_2 algorithm, then the label for the signal is automatically set to unrealiable.~~

~~To compare possible reliability labels, we train three initial and primitive classifiers using the previously engineered features. The first with a reliability label taken from agreement between the wrist worn device and the reflective fingertip sensor, the second taken between the wrist worn device and the fingertip transitive sensor, and the last classifier uses a level of agreement between all three sensors. The classifiers are all validated on unseen data.~~

Ground Truth Device(s)	Initial Weighted Precision Score
Fingertip Reflective	0.89
Fingertip Transitive	0.69
Both Ground Truth Devices	0.1

~~The high expectation of agreeance between 3 devices incurs too much noise for a classifier to predict. That is, it is too much for a classifier to compensate for errors across 2 ground truth devices.~~

~~The reflective sensor, and its algorithmic output, is chosen as the ground truth device. It is an obvious decision due to the finer granularity readings, lack of bias (since it is the same sensor), and higher initial precision in experiments~~.

C. Scoring

As discussed, to score the classifier in both training and validation we focus on average precision, weighted in certain experiments to compensate for the general imbalance between reliability labels. There are often a much larger number of unreliable readings than there are reliable. Precision is 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 are concerned with a high true positive score, and a low false positive, with little concern for false negatives. This holds as a reliable scoring metric as long as our true positive score provides frequent enough readings to build worthwhile applications.

D. Training

To select a classifier that best generalizes this approach to other data, we analyze several binary classifiers in the SciKit learn library and compare there results against multiple validation sets. Gradient boosting classifiers are found to provide robustness, generalizability, and the most consistently useable results. Despite SciKit-Learn having an implementation of this classifier, we use the python bindings offered with the XGBoost library [11], as it provides similar results for the same classifier with faster training times. We perform a cross validated grid search of hyperparameters using SciKit Learns[12] *GridSearchCV* to further tune the classifier. Optimal performance based on data trained across several individuals yields the following optimized parameter set. Data was trained using 5 fold cross validation, with shuffling prevented, overlaps between splits removed, and scoring tweaked due to the fact that time series data is inherently not independent and identically distributed. Individual folds were checked against a separate validation set.

Learning Rate	0.5
Number of Estimators	300
Booster	Tree
Objective	Binary Logistic

VI. Tuning

Outside of the parameters of the classifier itself, various hyperparameters within the problem domain are tuneable. The comparison of some of these are explored in this section.

A. Threshold

The first obvious tuneable hyperparameter is the threshold for which we can consider a reading to be reliable. Although a smaller threshold improves the confidence of our classifier, a larger threshold can will likely prove easier to predict. The threshold used will also be dependent on the applications being developed using the readings produced. An application that is looking at whether a users oxygen saturation level falls below what is considered healthy levels (90%) will have a higher tolerance for error than an application that expects near exact values.

Fig. 12. Precision of Classifiers vs Reliability Thresholds

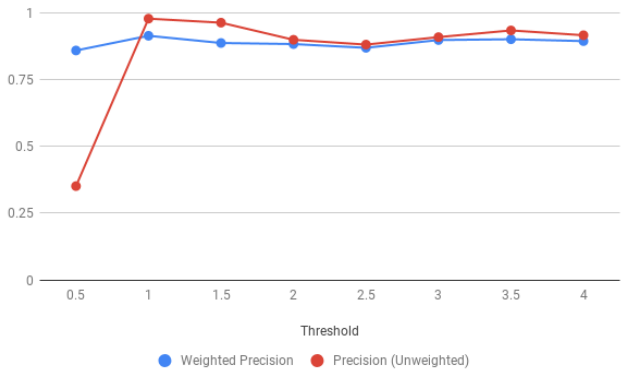
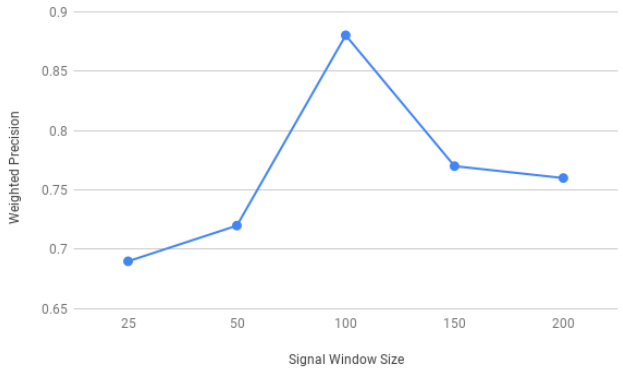


Figure 12 shows various thresholds against the precision as well as weighted precision to compensate for label imbalance that occurs in some experiments. Although thresholds of $\pm 1.0\%$ or greater provide satisfactory performance, it appears a threshold of $\pm 1.0\%$ provides the best results with this validation set.

1) Window Size: Although it is sensible to approach the problem using a window size equal to that of the algorithm that will be used to calculate the final SpO_2 reading, it is possible that more or less data will benefit the classifier. More data could potentially provide more context to the signal and hint at a malformed reading, and less data could potentially remove noise that the classifier would otherwise pickup. We verify these assumptions in figure 13

Figure 13 verifies that analyzing signal features on a window size equal to that of a signal used to calculate SpO_2 leads the highest precision on a validation set. A window size of 100 is used in the remained of the experiments.

Fig. 13. Precision of Classifiers vs Signal Window Sizes



2) Base Algorithm: We have analyzed two similar algorithms used to calculate SpO_2. Although intuitively, the *enhanced* algorithm should be used as a basis since its filtering is more robust and should aid the classifier, it is possible that this approach is pruning potentially reliable results. That is, it is possible the *enhanced* algorithm is too greedy in its consideration of unreliable values, and potentially ignoring valid signals. We compare the output of the reliability classifier used with each algorithm below.

TABLE I
COMPARING PRECISION OF CLASSIFIERS WITH DIFFERENT BASE ALGORITHMS

Base Algorithm	Precision Weighed
Enhanced	0.91
Maxim	0.61

Table I shows the weighed precision of a classifier trained with each algorithm used to generate the readings and subsequent reliability labels. Our original assumption holds in that the enhanced algorithm improves our reliability classifiers performance by filtering easily identifiable poor signals.

VII. VALIDATION

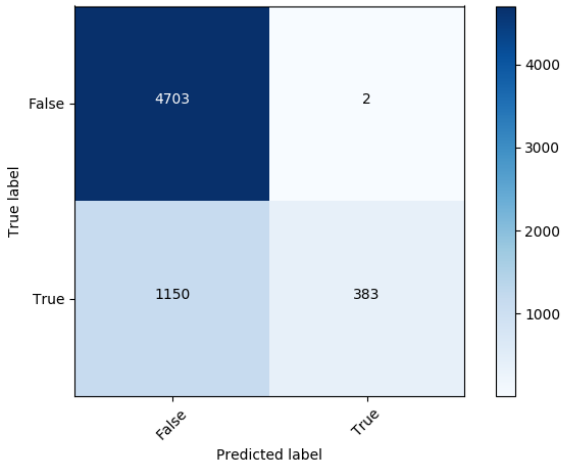
Figure 14 shows the confusion matrix of a model applied to data collected from a separate session than that with which it was trained. The model was trained on ten minutes of data and then applied to a 4 minute trace of unseen data from the same user in a separate session. Although transferability to new users is desirable, and discussed in the next section, it is useful to know that highly accurate results can be achieved for a specific user with small amounts of training data.

A. Transferability

1) User 1: To show that the classifiers precision is not restricted to a single person, tests were run on several other users who varied in weight, age, and skin colour.

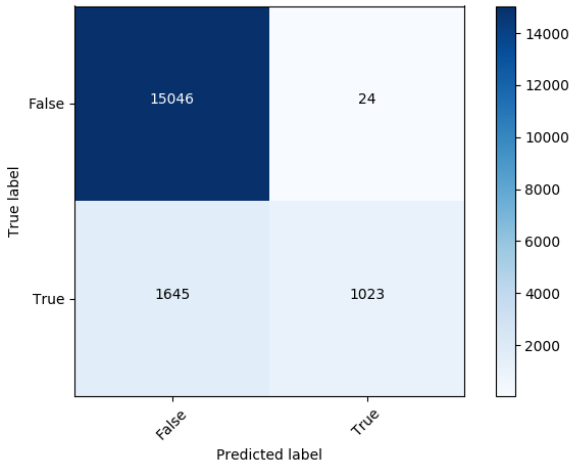
Figure 15 shows a confusion matrix with the results of the gradient boosting classifier applied to data from a person that it was not trained on. The classifier was trained on approximately 1/2 hour worth of data from a single person, consisting of approximately 70,000 instantaneous

Fig. 14. Model Applied to Unseen Data



device SpO_2 measurements. The validation set is taken from a different user and contains approximately 18000 labels. It is worth noting that the two individuals are of different genders, skin colours, and are approximately four years different in age. The confusion matrix corresponds to a weighted precision score of 89%.

Fig. 15. Confusion Matrix Unseen User 1



The red line in figure 16 shows the predicted values of the wrist worn device that are marked as useable by the enhanced SpO_2 algorithm. The black line in the figure shows the remaining values after the incoming signal are filtered by the reliability classifier. The blue line show the SpO_2 output of the fingertip sensor as a reference point.

As a comparison to the base algorithm, we calculate the mean squared error between the baseline enhanced algorithm and the fingertip readings, throwing away any readings where no result was returned by either. And then perform the same calculation with the output remaining after the reliability classifier prunes values. The baseline algorithm has a root mean squared error of 4%, and after values are pruned by the reliability classifier, it achieves a root mean squared error

Fig. 16. Readings of Enhanced Algorithm vs Reliability Classifier (User 1)

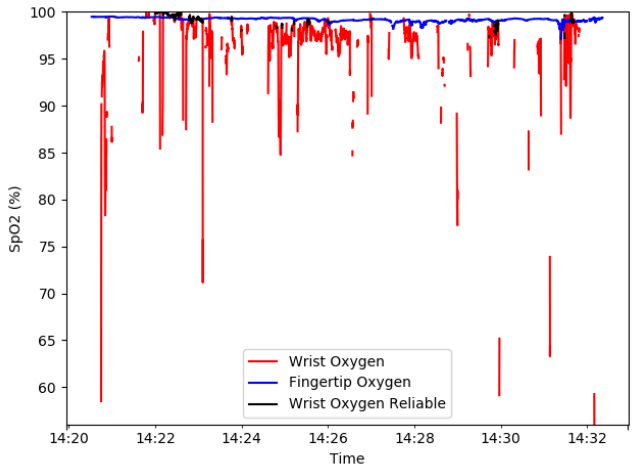
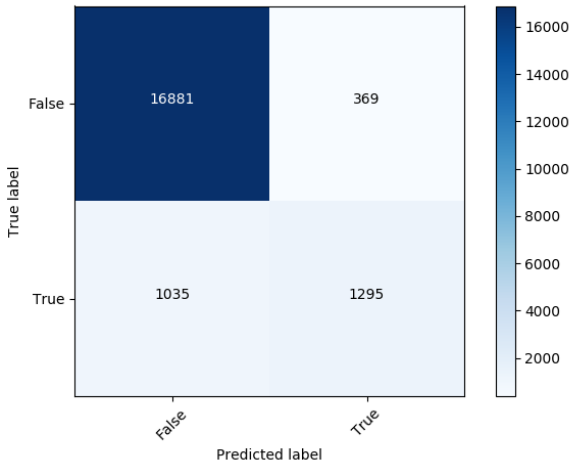


Fig. 17. Confusion Matrix Unseen User 2

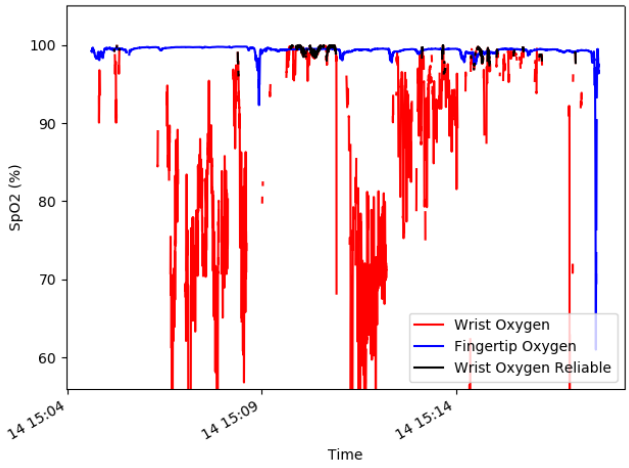


of 0.5%.

Despite that only 1023 readings remain after pruning (an 84% decrease in the number of reliable readings from the baseline algorithm), the largest time gap between any 2 reliable readings is only 2 minutes and 43 seconds. This frequency should be adequate to build a majority of useful applications, such as providing ample warning for patients potentially entering a state of hypoxemia.

2) User 2: We show similar measurements for a second unseen participant. The second unseen user is the same gender and of similar age to that of the training user, with significantly darker skin tone. Despite the potential issues with darker pigmentation referenced in section IX, the classifier performs well. Figure 17 shows the confusion matrix for a classifier trained on labels within $\pm 1.0\%$ of the wrist-worn device. The confusion matrix corresponds to a precision and weighted precision of 78% and 92% respectively. Figure 18 shows the same visualization of reliable readings, and in this case 1401 readings were marked reliable by the classifier compared to 8536 marked reliable by the baseline enhanced algorithm. Despite the higher count of false positives in this users results, the root mean squared error still dropped from 19.4% in the baseline enhanced algorithm compared to 0.64% after pruning signals marked unreliable by the classifier. The greatest distance between any two reliable readings for this user was only 3 minutes and 1 second.

Fig. 18. Readings of Enhanced Algorithm vs Reliability Classifier (User 2)



B. Differences Across Trials

Users do not always reliably collect data the same across different trials. This sections describes a second trial run on the user described in section VII-A.2. Setting the threshold of the reliability classifier to $\pm 1.0\%$ for this user in the second trial yields zero useable readings. After analyzing the data, we can see that the only 0.2% of the readings produced are within a threshold of $\pm 1.0\%$ from the fingertip sensors output. That is, there are few reliable results for the classifier to mark, and in this case the classifier is unable to find them.

In a second analysis, we raise the threshold to $\pm 3.0\%$ and rerun the classifier on this user. The output of the baseline algorithm against the reliability classifier is shown in figure 19.

Although only 1.1% of the labels are truly within $\pm 3.0\%$ of the wrist-worn devices output, the reliability classifier marks 3 times as many labels as reliable. Although this should indicate a high rate of error, the root mean squared error was still reduced from 14% to 5% between the baseline algorithm and the pruned results after classification. This is indicating, that although the values being misclassified as reliable by the classifier are outside of the threshold set, they are not far away from the threshold range and still potentially useable.

This is inherently a problem of using binary classification, although from validation we know a label is wrong, we have no way of knowing how wrong. One solution to this would be to extend this method to utilize multi-label classification, indicating a level of confidence that each signal will produce a label that resides in one of several threshold ranges from the ground truth. We discuss this approach further in section

Fig. 19. Readings of Enhanced Algorithm vs Reliability Classifier (User 3)

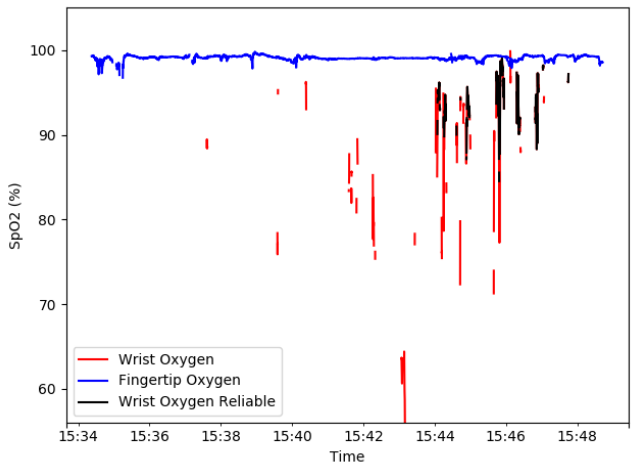
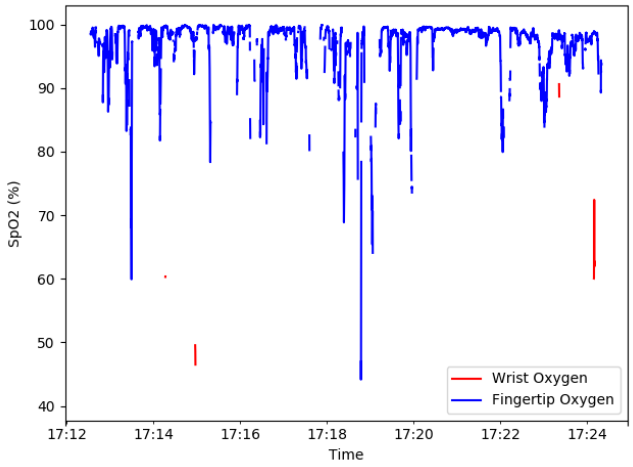


Fig. 20. Poor Data Collection from User



VIII-C. This would also aid in increasing the frequency for which reliable signals are produced, as the user during this trial experiences a nine minute delay where no reliable signal is marked by the classifier.

Increasing the threshold of the classifier applied to this trial even further to $\pm 4.0\%$, maintains the same 5% root mean squared error, increases the number of classified reliable signals to 321, and reduces the delay between reliable signals from 9 minutes to a 3 minute frequency.

C. Difficulty in Data Collection

It should be noted that occasionally a trial on a user would produce no useable labels for a classifier to accurately predict, except with unreasonably large thresholds. For one user in particular, out of 12 minutes of data, or approximately 18000 labels, only 37 signal windows were marked reliable by the baseline algorithm. Out of those 37 signals, the closest one to a ground truth reading from the wrist-device showed a 10% error. Figure 20 shows the trace of SpO_2 collected from the wrist and fingertip of this user. This data shows significant error in both the wrist-worn device as well as the fingertip attached sensor. One of two possible reasons for this error exist, either the devices were attached poorly to this user, or some trait of the user makes them a poor responder to data collection by this method. This user was the same age, gender, and similar skin tone to at least one other participant, indicating that poor device attachment was the most likely cause. It should be noted that reliability classification applied to this trace marks all 37 wrist-worn device readings as unreliable, which is the desirable result given that the most accurate value has a 10% error.

VIII. Discussion

This section covers some future work for the project, including necessary steps to deploy the method in the wild on consumer grade devices.

A. Reducing Compute Costs

As we've shown, this method can improve the confidence in how reliable a signal is when used for SpO_2 measurements. Currently this approach comes with a high computation cost, for this method as it stands 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. Firstly, 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 pruning features, work similar to Sidewinder[13] could be used to offload signal reliability calculation to a lower powered processor, and subsequently wake the device when a useable 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 transfer time for up to one hours worth of data can be performed within seconds to a remote server, and could easily be run on a device wake up so as to not effect mobile battery and performance.

Currently, performing a full tsfresh feature extraction for approximately half of an hour of data takes roughly 1 hour on 8 Intel Xeon E5-2680 CPUs. Although the time to execute this process can immediately be reduced by half by making the process of pruning features from the classifier cyclic. That is, once zero-significance features are pruned by the classifier, use those results as the feature selection input to tsfresh on future runs. Currently we do not pass these features back, which is to ensure experiments are independent.

B. Improved Ground Truth

It is possible that the confidence in each threshold could be increased by utilizing a more reliable ground truth. If a transitive sensor provided a fine enough granularity and had passed a rigorous medical calibration, it could be used as a more trustworthy ground truth for creating reliability

labels. A necessary feature of the transitive sensor that we employed was the ability to access the raw readings as they are produced, which allows us to attach a timestamp to the reading for direct comparison to the other devices. Although we had multiple transitive pulse oximeters in our possession, the one employed in the experiments was the only one with which we were able to reverse engineer the USB serial protocol to access the live streamed data. In future experiments it would be desireable to research a larger plethora of transitive pulse oximeters to find a medical grade device that allows lower level access to the readings produced.

C. Improving Classification

Although there is a variance in the people that the classifier is trained, tested, and validated on, the small subset of people used in the study (5 across all experiments) could be limiting the ability of the classifier to predict some values if they are all within a healthy range. Future work will include a larger subset of participants, if possible incorporating some people with symptoms contributing to a lowered average SpO_2. Training with more people would also require more aggressive feature pruning discussed in section VIII-A to make testing iteration times reasonable.

Finally, a major improvement on the current classifier would be extending the output to include multi-label classification. 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.

D. Extending to Existing Mobile Devices

The next obvious iteration of the wrist-worn device is to either build or utilize a consumer grade device. If low level access to the LED reflectance and IMU motion traces were provided for an existing consumer grade device, the same approach could be applied to existing hardware to ensure none of the approaches taken are influenced by the device itself. The use of existing device could potentially improve results solely based on the quality of the hardware. Although as discussed in section III-B.1, the methods could potentially need adjustment to compensate for things such as additional ambient light. Despite this, changes should be limited to classifier retraining since feature extraction is near exhaustive and automated.

IX. Related Work

Ra et al. perform similar analysis and reliability detection to our work in the context of heart-rate measurement [14]. A majority of work done in quantifying pulse oximetry pertains to its use of heart rate measurement, such as rule based detection of heart rate for reliability [15]. 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][16]. Possible wearability sites, including the wrist, and various sensor configurations have been considered in the context of telehealth monitoring [17][18]. Other work has documented the process of building not only devices, but working 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 increase 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. Yan et al.[26] used a more sophisticated feature extraction to remove motion and other noise artifacts in the context of at home 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.

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.

X. Conclusion

In this work we create a test bed to collect several signal channels from a wrist-worn device and compare their algorithmic output for calculating SpO_2 against more dependable devices, and the same devices in more dependable configurations. We implement a pipeline for the analysis and visualization of this data offline. We demonstrate the difficulty that existing algorithms have with producing an oxygen saturation reading from wrist-worn devices, that can match up with readings from more reliable measurement sites and sensors. We use automated feature extraction on various user traces to build a classifier that successfully predicts highly reliable SpO_2 output solely from the signals given by sensors common to many wrist-worn mobile devices. We show that the approach generalizes to unseen data from the same user, and generalizes to new users with highly varied experimental conditions. Although work needs to be done to deploy this in the wild, we believe that this holds as an adequate proof of concept for a viable signal pruning and filtering technique for SpO_2 values measured from wrist-worn devices.

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," *European Respiratory Journal*, 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: The Journal of the American Society of Anesthesiologists*, vol. 112, no. 2, pp. 282–287, 2010.

[4] Y. Mendelson and B. D. Ochs, "Noninvasive pulse oximetry utilizing skin reflectance photoplethysmography," *IEEE Transactions on Biomedical Engineering*, vol. 35, no. 10, pp. 798–805, 1988.

[5] J. E. Sinex, "Pulse oximetry: principles and limitations," *The American journal of emergency medicine*, vol. 17, no. 1, pp. 59–66, 1999.

[6] 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 *Communication and Signal Processing (ICCSP), 2016 International Conference on*, pp. 1136–1140, IEEE, 2016.

[7] M. Integrated, "MAX30102 high-sensitivity pulse oximeter and heart-rate sensor for wearable health," 2018.

[8] "Usb pulse meter bm3000b http://www.shberrymed.com/usb-pulse-meter-bm3000b-p00037p1.html."

[9] MolecularD, "Pulse oximeter with much improved precision: 6 steps (with pictures)." `https://www.instructables.com/id/Pulse-Oximeter-With-Much-Improved-Precision/`. (Accessed on 08/26/2018).

[10] 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.

[11] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, (New York, NY, USA), pp. 785–794, ACM, 2016.

[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*, vol. 12, pp. 2825–2830, 2011.

[13] 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 Computer Architecture News*, vol. 44, no. 2, pp. 205–215, 2016.

[14] 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 *Proceedings of the 18th International Workshop on Mobile Computing Systems and Applications*, HotMobile '17, (New York, NY, USA), pp. 49–54, ACM, 2017.

[15] 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.

[16] 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 *Engineering in Medicine and Biology Society, 2005. IEEE-EMBS 2005. 27th Annual International Conference of the*, pp. 3585–3588, IEEE, 2005.

[17] 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*, vol. 4, pp. 3016–3019, IEEE, 2003.

[18] Y. Mendelson, R. 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*, pp. 912–915, IEEE, 2006.

[19] S. Bagha and L. Shaw, "A real time analysis of ppg signal for measurement of spo2 and pulse rate," *International journal of computer applications*, vol. 36, no. 11, pp. 45–50, 2011.

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

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

[22] J. Emery, "Skin pigmentation as an influence on the accuracy of pulse oximetry.," *Journal of perinatology: official journal of the California Perinatal Association*, vol. 7, no. 4, pp. 329–330, 1987.

[23] 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.

[24] 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.

[25] J. Yao and S. 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*, vol. 1, pp. 2153–2156, IEEE, 2004.

[26] Y.-S. Yan and Y.-T. Zhang, "An efficient motion-resistant method for wearable pulse oximeter," *IEEE Transactions on information technology in biomedicine*, vol. 12, no. 3, pp. 399–405, 2008.

[27] 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 *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services Companion*, MobiSys '16 Companion, (New York, NY, USA), pp. 139–139, ACM, 2016.