Evaluation of the Signal Quality of Wrist-Based Photoplethysmography

Nikhilesh Pradhan, Sreeraman Rajan and Andy Adler

Systems and Computer Engineering, Carleton University, Ottawa, Canada

E-mail: Andy.Adler@Carleton.ca

Abstract. *Objective:* Wearable devices with embedded photoplethysmography (PPG) sensors enable continuous monitoring of cardiovascular activity, allowing for the detection cardiovascular problems, such as arrhythmias. However, the quality of wrist-based PPG is highly variable, and is subject to artifacts from motion and other interferences. The goal of this paper is to evaluate the signal quality obtained from wrist-based PPG when used in an ambulatory setting. *Approach:* Ambulatory data were collected over a 24-hour period for 10 elderly, and 16 non-elderly participants. Visual assessment is used as the gold standard for PPG signal quality, with inter-rater agreement evaluated using Fleiss' Kappa. With this gold standard, 5 classifiers were evaluated using a modified 13-fold cross-validation approach. *Main results:* A *Random Forest* quality classification algorithm showed the best performance, with an accuracy of 74.5%, and was then used to evaluate 24-hour long ambulatory wrist-based PPG measurements. *Significance:* In general, data quality was high at night, and low during the day. Our results suggest wrist-based PPG may be best for continuous cardiovascular monitoring applications during the night, but less useful during the day unless methods can be identified to improve low quality signal segments.

1. Introduction

Long-term, continuous, non-invasive monitoring of cardiovascular activity in an ambulatory setting would enable early detection of heart diseases such as arrhythmias, thereby allowing for early interventions to help reduce emergencies such as strokes and heart attacks.

However, prolonged, ubiquitous monitoring is impractical with electrocardiogram (ECG) Holter monitors, the technology currently employed medically for ambulatory use. ECG Holter monitors are cumbersome and uncomfortable due to their requirement for electrodes and wires adhering to the thorax, and also place restrictions on the user. Since the monitors are sensitive to water damage, users are unable to shower or swim. As a result, users are less likely to comply with instructions to wear Holter monitors continuously for prolonged periods, especially if they are not currently suffering from cardiovascular problems.

A potential alternative to ECG Holter monitors is wrist-based photoplethysmography (PPG), an optical technology that is easily embedded into wearable devices such

as wristbands, and is already included in many commercially available smartwatches. PPG detects local changes in blood volume by measuring the attenuation of light. PPG waveforms consist of both a varying "AC", and a large, constant "DC" component. The DC component accounts for the majority of the attenuation, from the absorption of light by the skin, bone, venous blood, and other tissues. The AC fluctuations are driven by the cardiac cycle.

Unfortunately, the PPG signal is highly sensitive to noise-corruption, especially due to motion (Allen 2007, Kamal *et al* 1989, Abdallah and Bolz 2011). Cardiovascular parameters derived from noise-corrupted data are unreliable and could lead to inaccurate diagnoses. This necessitates the assessment of signal quality, to ensure that noise-corrupted signals are handled appropriately — either cleaned, or rejected prior to cardiovascular analysis.

The signal quality algorithms employed by smartwatch vendors are proprietary, and largely unpublished. Current PPG-analysis algorithms in smartwatches are designed primarily for heart rate (HR) monitoring, which can use frequency-based algorithms and do not require beat-to-beat analysis. This means that the vendor quality algorithms correspond to different criteria to those for cardiovascular monitoring. Li and Clifford (2012) proposed a template matching scheme using dynamic time warping to assess quality. However, it relies on accurate beat detection for proper assessment. While this approach may be useful for sedentary in-patient clinical finger-clip PPG signals, beat detection is difficult when the data has large artifacts, as is the case for wristband based PPG data collected from prolonged monitoring of mobile individuals. Morris and Wander (2014) used a method combining segmenting and non-segmenting approaches to derive features for classification. Their study used PPG data with motion artifacts simulated by requiring the participants to perform certain behavioural tasks. Sukor *et al* (2011) proposed a method for the detection of noise-corrupted heart beats in the PPG signal using morphological features of the waveform, such as pulse amplitude, trough depth difference, and pulse width. Subjects were asked to perform eight different hand movements to simulate a variety of motion artifacts, though motions artifacts from the finger-clip PPG sensor system may not be consistent with those present in wrist-based PPG systems.

To understand the quality during prolonged monitoring, we identify several questions which have not been directly answered in the literature. Previous published works used short-time lab samples rather than 24-hour, ambulatory, wrist-based PPG data for the development and evaluation of their proposed algorithms. Furthermore, the gold standard for signal quality ratings, which is visual assessment, is susceptible to subjective bias, thus necessitating a larger number of raters. Quantitative data regarding the quality of ambulatory wrist-based PPG is also limited, despite being valuable for designers intending to use the technology in various continuous monitoring applications. Some of our early results were published as McCarthy *et al* (2016), where we found that 45% of the acquired signal from wrist-based PPG during a 24-hour period was of high quality. However, these preliminary results relied on visual assessment of the PPG

signal quality, thus it was only practical to assess the quality of short samples of data throughout the 24-hour period

2. Experimental Methodology

Simultaneous wrist-based photoplethysmography (PPG) and electrocardiogram (ECG) data was collected from 26 healthy participants for a period of 24 hours as they performed their daily routine. The study was approved by the Carleton University Research Ethics Board.

2.1. Participants

Participants were recruited from two broad subject sets: 1) elderly, defined as individuals 65 years of age or older, and 2) non-elderly, defined as anyone not included in the elderly group. This enabled the evaluation of differences in the quality of wrist-based PPG between the two groups, as the groups were expected to have different types of activities and schedules.

For participation in this study, subjects were required to be healthy and mobile, defined as having the ability to walk without requiring assistive devices. This requirement was used as motion artifacts during daily activities were expected to be the primary source of signal artifacts. The daily routines of participants with compromised mobility would be expected to include considerably less motion than the average person.

Ten elderly participants were recruited through advertisements in relevant newsletters and through contacts, and sixteen non-elderly participants were recruited from university students and researchers. Further demographic information such as age, sex and ethnicity were not collected.

2.2. Devices

PPG signals were collected using an Empatica E4 wristband (Empatica, Milan, Italy), while ECG signals were collected using a Seer Light Extend Holter monitor (General Electric Healthcare, Chicago, USA). The Empatica E4 is a wrist worn device, similar to a smartwatch, equipped with other sensors such as a temperature sensor (infrared thermopile), a 3-axis accelerometer, and a skin conductance sensor, of which the PPG and accelerometer signals were used in this study. During the study, signals were recorded continuously and stored on the device, with she PPG signal sampled at 64 Hz, and the 3-axis accelerometer signal is sampled at 32 Hz‡. Simultaneously, the Holter monitor was used to obtain the ECG signals; however, this data was not used for this study.

‡ Empatica, E4 Wristband User Manual, 2015. https://empatica.app.box.com/v/E4-User-Manual

2.3. Experimental Setup

Participants were asked to wear the Empatica E4 wristband and the GE Seer Light Extend Holter monitor. The Empatica E4 wristband was worn as a wristwatch, with the sensor facing the posterior side of the participant's wrist on their non-dominant arm. The non-dominant arm was expected to engage in less movement relative to the dominant arm, potentially providing superior signal quality.

For a period of 24 hours, the participants wore both devices. As the Holter monitor was susceptible to water damage, participants were asked not to swim or bathe. With this exception, they were asked to continue with their daily routine as much as possible.

After the data collection period, the devices were removed and the data extracted. The devices were cleaned with alcohol prior to use for the next participant.

2.4. Limitations of Experimental Setup

Certain limitations occurred due to scheduling. Occasionally, participants were unable to meet with the researchers at the scheduled times, and thus, the data collection period was sometimes slightly shorter than 24 hours. In other cases, participants were unable to meet with the researchers the day following the device setup, and were therefore instructed to remove and power off the devices themselves.

The Empatica E4 wristband was worn as tightly as the participants felt was comfortable. Since the tightness was based on acceptability by the participant, some participants wore the devices more loosely than others. Those that wore the device more loosely are likely to provide inferior signal quality, as the wristband had a greater freedom of movement, and motion is expected to be a cause of noise corruption in the signal.

Participants were not required to keep an activity log for the study, therefore, their sleep and exercise times are not known. This information may have enabled additional analysis and understanding of the data. Participant compliance during the data collection period was not monitored.

3. Development of a gold standard classification

Establishment of the gold standard against which to compare automatically detected features was done by selecting a subset of data and requesting manual classification by raters based on visual assessment, which can be subjective to the raters. To better understand the variability of ratings due to subjectivity of raters, our gold standard was created using a compilation of ratings from 17 raters. Statistics were then computed to quantify the extent of classification agreement between raters. Previous work relied on 2–3 raters, and did not disclose the extent of agreement between those raters (Wander and Morris 2014, Li and Clifford 2012, Sukor *et al* 2011). Thus the reliability of the gold standards used in these works cannot be assessed.

PPG ratings were established as follows: the entire data set for each participant was subdivided into 10-second segments. A random number generator was used to select 39 of the segments for analysis from each of the 26 participants, resulting in a set of 1014 non-overlapping, non-continuous PPG segments, to be used for analysis. As the segments were chosen randomly, there were no controls to ensure equal quality-class representation. Random selection was chosen, as there were no pre-existing classifications that could be used to ensure equal class representation in the set. It is expected that the random selection process would result in a class representation roughly proportional to the class representation in the overall data.

Each 10-second segment was assigned a signal quality levels from 1 to 5, with 1 representing the lowest quality, and 5 representing the highest quality. The quality levels were defined by the percentage of the segment for which clear pulses with discernible peaks were identifiable. A class 5 segment must have all identifiable pulses for the entire data segment, class 4 for at least 75% of the segment, class 3 for at least 50% of the segment, class 2 for at least 25% of the segment, and class 1 for less than 25% of the segment (figure 1). These criteria were motivated by the quality level definitions used by Wander and Morris (2014), who used the number of peaks visible within the data segment. A high quality PPG signal is expected to contain a series of pulses of approximately uniform morphology, as shown in the class 5 quality level in figure 1.

A graphical user interface was created in Matlab for the raters to annotate the PPG segments. The 17 raters were recruited from biomedical engineering students at Carleton University. Each rater was provided with the definition of the quality classes, as well as two examples of each class, as identified by the researchers. The raters could complete the annotations over multiple sittings.

3.1. Agreement between Raters

Annotations made by the raters were compiled, and for each PPG segment, the class chosen by a plurality of raters was used as the gold standard class for the segment. Agreement between raters, P_i, was computed for each PPG segment using (1), the Fleiss (1971) κ. According to this measure, $P_i = 1$ indicates that all raters agreed on the classification of the segment. The P_i values from all the PPG segments were collated together to produce the histogram in figure 2.

$$P_i = \frac{1}{n(n-1)}\Big(\sum_{j=1}^{k} {n_{ij}}^2 - n\Big) \tag{1}$$

where n is the total number of raters (in this case 17), k is the number of quality classes (5 in this case), i is the segment number, and n_{ij} is the number of raters who classified segment i to class j.

Only 11% of the PPG segments had perfect agreement between all 17 raters. Thus the majority of PPG segments had an element of subjectivity, despite a clear set of rules governing class membership. Approximately 58% of the segments had a $P_i > 0.5$, with the remaining 42% showing poor agreement between raters.

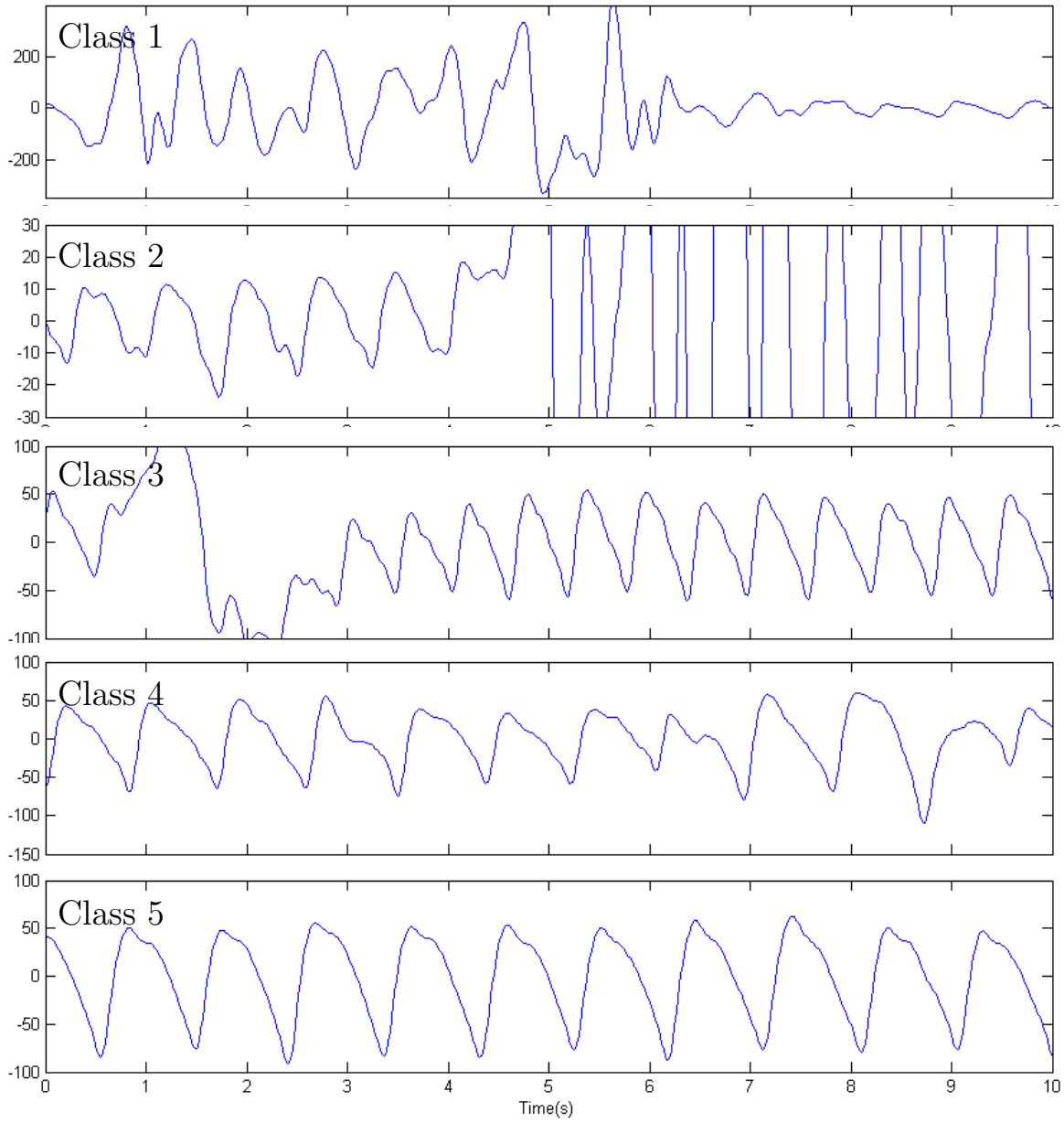

Figure 1. Examples of PPG Signal belonging to class 1 (top) to class 5 (bottom). A class 5 segment must have all identifiable pulses for the entire data segment, class 4 for at least 75% of the segment, class 3 for at least 50% of the segment, class 2 for at least 25% of the segment, and class 1 for less than 25% of the segment.

To obtain a normalized measure of the strength of agreement between raters, Fleiss' Kappa, κ, was computed. The agreements (P_i) computed for each segment were averaged to obtain $\bar{P}$, referred to as the extent of agreement. This value may be compared to the expected extent of agreement was calculated if ratings were made randomly, $\bar{P}_e = \sum_{j=1}^{k} P_j^2$. Fleiss's Kappa, κ (2), is computed as the ration of the total possible extent of agreement beyond the agreement due to random chance (1-$\bar{P}_e$) and

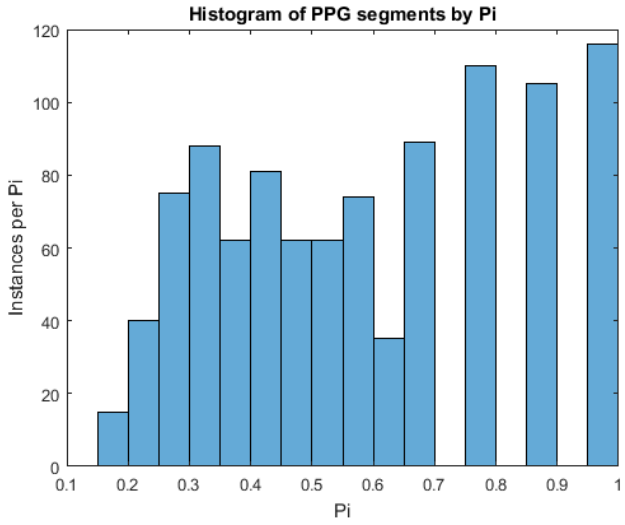

Figure 2. Histogram showing the distribution of PPG segments used for analysis, organized by the agreement between raters per segment, P_i.

the agreement obtained in excess to random chance ($\bar{P} - \bar{P}_e$).

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} \tag{2}$$

For this study, we obtained, $\kappa = 0.4605$, which, according to the benchmarks established by Landis and Koch, 1977), indicates moderate agreement between raters.

4. Feature Selection

Features were extracted from each of the 1014 ten-second data segments. Some features were adopted from metrics recommended in previous literature, while others were developed based on observed differences between the data belonging to the different classes. A complete discussion of all 71 evaluated features is given by Pradhan (2017), while this paper described the features selected for best found quality classification.

A reduction in the number of features was desired for efficiency, by eliminating poor and unnecessary features. However, the univariate analysis of the features indicated that no individual feature was adequate for discriminating between the five signal quality classes. Therefore, a multivariate approach was used to identify a subset of features, which together would provide the best class discrimination.

Using Weka (Frank *et al* 2016), feature selection was done with the wrapper method, which performs various iterations with different subsets of features, and selects the subset offering the maximum class discrimination. The wrapper method is preferred to ranker methods as it considers feature dependencies and correlations for feature subset selection (Guyon *et al* 2003). Feature dependency indicates whether multiple features together can provide good class discrimination where each feature individually would not. Under an individual ranker approach, such dependencies would not be considered in the feature

assessment (Saeys *et al* 2007). Feature correlation identifies features which may provide similar class discrimination information, thus making them redundant. Therefore, a wrapper method was chosen for the feature selection process.

Wrapper methods evaluate subsets using a classifier, providing a direct link between the feature selection and the classifiers used in the classification stage (Guyon *et al* 2003). We used a random forest classifier, as it was deemed to be sufficiently complex to provide good results for the features with poor univariate discriminability, while being simple enough not to be computationally intensive or time consuming.

To obtain results in a reasonable time, a greedy step-wise approach was used. While an exhaustive method is generally considered to be superior, it is computationally intensive and therefore time consuming for searches with a high number of features, as in this case. A drawback of the greedy step-wise approach is its susceptibility to getting stuck in a local optimum, rather than finding the overall optimum as an exhaustive search would (Saeys *et al* 2007). To mitigate this, both forward selection and backward elimination greedy step-wise methods were applied. Backward elimination begins with the full set of features, and attempts to remove individual features without compromising the discriminability of the feature subset, whereas forward selection begins with no features and works in reverse to identify the optimal subset. The performance of a particular subset is evaluated by Weka using a *merit score*, in a range from 0 (worst) to 1 (best), based on the classification accuracy. Backward elimination resulted in a subset of 70 features, with a merit score of 0.778. Forward selection resulted in 9 features, with a merit score of 0.777. Due to the similarity in merit score for both methods, the 9 feature subset from the forward selection search was selected. Repetition of the forward selection search can be used to strengthen the confidence in the results if repetitions produce the same subset of features. Forward selection was repeated 3 times, with the same 9 features being selected at each repetition. The selected features are as follows:

- *Number of Peaks by Billauer's Algorithm:* This feature is the number of peaks in the PPG signal identified using the algorithm of Billauer (2012) for peak detection. Rather than using a derivative-based approach, the algorithm identifies peaks based on preceding values being lower than a threshold.
  The algorithm searches for a peak by traversing the data segment point-by-point while tracking the highest point encountered, until it reaches a data point that is lower by a specified threshold. Once such a data point is reached, the current highest point is recorded as a peak value, and the current data point recorded as the lowest point. Then the algorithm continues traversing the data segment searching for a trough (a valley point) in the converse way. This process is repeated as the algorithm traverses the data segment.
  This was implemented in Matlab using the *peakdet* function from Billauer (2012). As seen in figure 3, a greater number of peaks were expected to be found in data belonging to lower quality classes.
- *Number of Zero-Crossings:* The Zero-Crossings feature is the number of times the

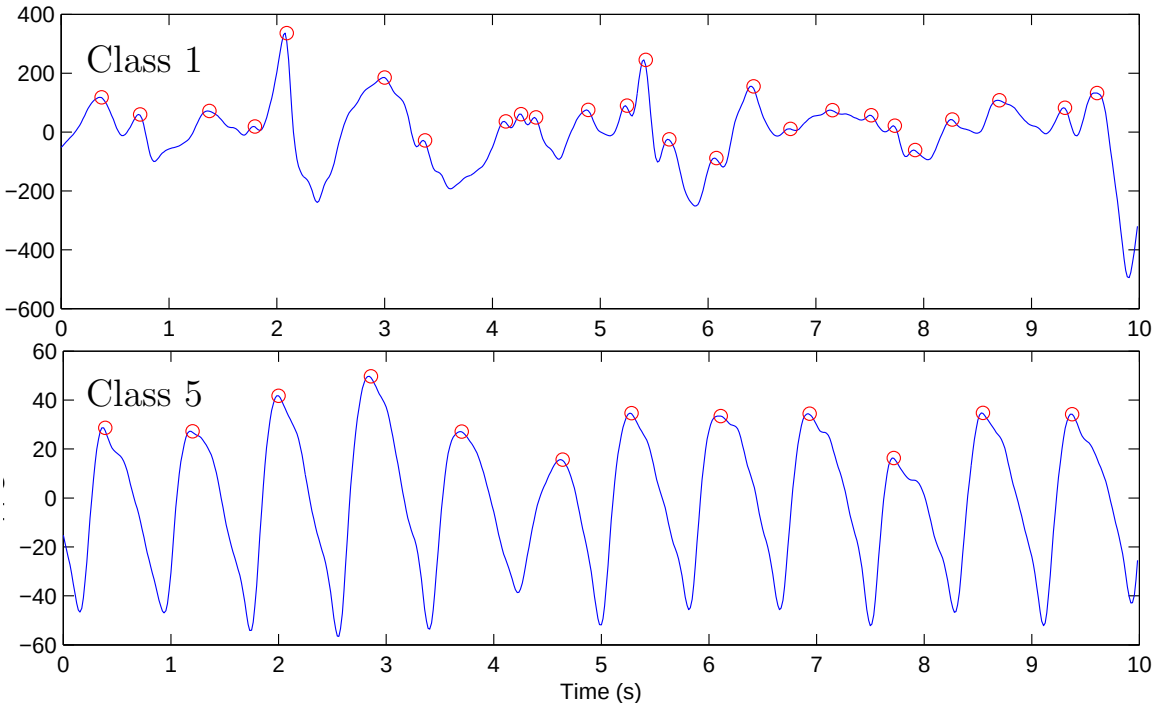

Figure 3. Example of peaks found using Billauer's algorithm in a class 1 and class 5 PPG segments.

PPG signal crosses the x-axis ($y = 0$), divided by the length of the data segment. It was expected that data segments belonging to lower classes would have a higher zero-crossing rate. This feature was inspired by the previous work of Elgendi (2016). Mean subtraction was not applied as the Empatica E4 wristband's internal computations remove baseline drift.

- *Accelerometer Features:* Empatica E4 wristbands contain 3-axis accelerometers that record data at 32Hz, continuously with the PPG sensor. Low quality data is expected to be found at times when the participant engages in arm motion. Such motion would register on the accelerometer in the axes corresponding to the direction of motion. The magnitude of the acceleration is computed to combine the x, y, and z axes of the accelerometer data. When the magnitude of the acceleration differs from the acceleration due to gravity, it is likely due to movement; hence the accelerometer data ought to correlate with PPG signal quality. This is seen in figure 4 (accelerometer signals), where the class 1 accelerometer signal has a larger range and different shape than the class 5 signal. These properties were represented using the median and the standard deviation of the accelerometer magnitude as features. The median accelerometer is an indication of baseline deviations from the acceleration due to gravity, while the standard deviation of the accelerometer magnitudes is an indicator of the range of the data, and the extent to which the data points are spread about the mean.
- *Correlogram Features:* It was postulated that there would be a difference in the

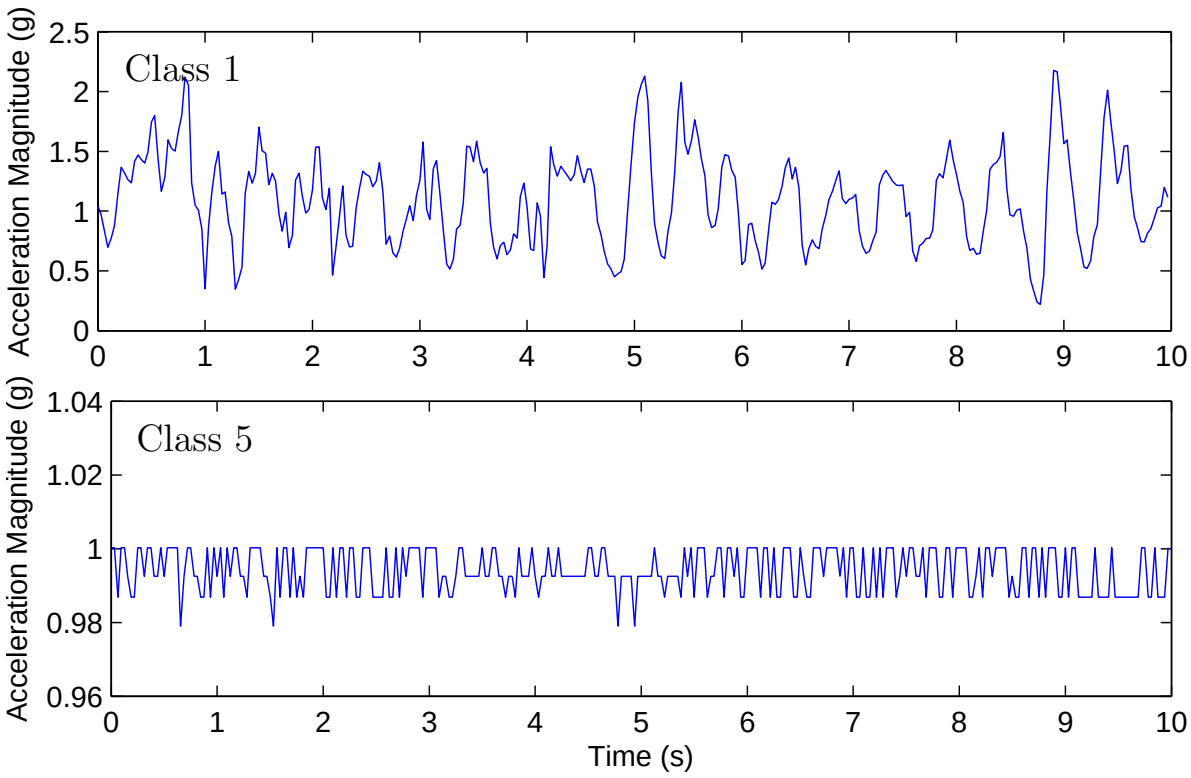

Figure 4. Magnitude of accelerometer signal for class 1 and class 5 data segments.

periodicity of the PPG signals of different classes. Noise corruption due to random, non-periodic motions were expected to result in non-periodic motion artifacts in the PPG signal. Thus, an assessment of the periodicity of the PPG signal may assist in class discrimination. This was done using a correlogram.

A correlogram displays the normalized autocorrelation of the PPG signal on the y-axis, with corresponding time lags on the x-axis. Autocorrelation values could range from +1.0, indicating perfect correlation, to zero, indicating no correlation, to -1.0, indicating a perfect inverse correlation. Correlograms were constructed for each data segment. The autocorrelation of the PPG data was determined at lags of up to 3 seconds.

In the correlograms for each data segment, the first two peaks were identified using the *findpeaks* function in Matlab. Peaks occur in the correlogram as the autocorrelation rises, associated with offsets at multiples of the period of the PPG signal. The features used were the values at the first two peaks in the correlogram. These features were inspired by Wander and Morris (2014), and previously used in preliminary work done by our research team (Pradhan 2017).

- *Median Noise Ratio per Pulse:* This feature is the median of the signal-to-noise ratio (SNR) computed from each pulse in the PPG segment. For each pulse in the ten-second PPG segment, (3) is applied to compute an SNR. This is an implementation of SNR by Elgendi (2016). In this implementation, SNR is defined as the ratio of signal variance to the noise variance. The signal variance is defined as the variance

of the absolute value of the PPG, while the noise variance is defined as the variance of the PPG. The median of the values of the NoiseRatio obtained from each each pulse in the ten-second segment is used as the feature.

$$NoiseRatio = \frac{\sigma_y^2}{\sigma_{|y|}^2} \tag{3}$$

where y is the PPG signal.

- *Median Relative Power per Pulse:* This feature is the median of the relative power computed from each pulse in the PPG segment. This is the ratio between the power of the Welch periodogram in the frequency range associated with good signals, 1–2.25 Hz, and the frequency range from 0–8 Hz, computed for each pulse, then the median of those values is taken as a feature. The relative power was computed according to (4). This feature was inspired from work done by Elgendi (2016).

$$RelativePower = \frac{\int_1^{2.25} P(f)df}{\int_0^8 P(f)df} \tag{4}$$

where $P(f)$ is the power spectral density at frequency f. Using sampled data the integrals were approximated by sums.

- *Standard Deviation of Shannon Energy Per Pulse:* This feature is the standard deviation of the Shannon energy computed from each pulse in the PPG segment (Liang *et al* 1997). It was computed using (5), which is the implementation used by Elgendi (2016) and Coifman and Wickerhauser (1992). The formula was erroneously referred to as entropy in the 2016 paper. We follow Elgendi (2016) and calculate the Shannon energy without normalization.

$$ShannonEnergy = -\sum_{n=1}^{N} \Big((x[n]^2) \Big(log_e(x[n]^2) \Big) \Big) \tag{5}$$

5. Classifier Selection

Five classifiers were evaluated using the annotated dataset and the nine features chosen from the feature selection process. Evaluations of each of the classifiers were done using a modified 13-fold cross validation training/testing scheme. This was done to maximize use of the annotated dataset, while ensuring that the classifier is not overfitting to the dataset. In 13-fold cross-validation, the dataset is divided into 13 subsets. At each iteration of the cross validation process, one subset is used as a testing set, while the remaining 12 subsets were used to train the classifier. This is repeated 13 times until classifier predictions have been made for each subset. Standard practice for cross-validation is to randomize the selection of subset membership, while ensuring class balance. However, due to concerns regarding overfitting, this process is modified. Instead, subset membership was determined manually to ensure that data segments from any one participant were not spread across multiple subsets. This ensures that all 39 data segments from each participant were co-located in the same subset. Thus,

at any given iteration, the classifier is not training and testing on data from the same participant.

Evaluation of the classifiers yielded various accuracies, depicted in table 1. The highest accuracy, 74.5%, was obtained from the Random Forest classifier, hence it was selected for the final algorithm. While this is the best accuracy obtained from the classifier analysis, it performs poorly for certain classes. However, the classes for which the classifier performed poorly appear to have low prevalence, thereby allowing an accuracy of 74.5%.

Random Forest is a meta-learning based approach using multiple decision tree classifiers. Each tree is trained using a randomly selected subset of instances and features. This allows each tree to specialise in discriminating between different subsets of data. The final classification for an instance (a 10-second segment) is obtained through a plurality voting system, in which each of the decision trees cast a vote. Thus, the weaknesses of one decision tree can be compensated by the others in the forest. This meta-learning approach is likely responsible for the higher accuracy obtained by this classifier.

Classifier	**Accuracy**
k-Nearest Neighbour	42.9%
Multi-Class SVM	43.5%
Naïve Bayes	63.6%
Decision Tree	66.9%
Random Forest	**74.5%**

Table 1. Accuracy of classifiers evaluated

6. Signal Quality over 24-hour Period

The *Random Forest* classifier was selected as the best choice for the signal quality algorithm based on the work done in the previous section. The classifier was then used to provide signal quality analysis for the full 24 hours of data collected from each participant.

6.1. Quality Assessment Procedure

For classifier selection, a modified 13-fold cross-validation scheme was used for the purpose of evaluating the classifier performance without overfitting. However, for this 24-hour assessment, accuracy of the results for this dataset were prioritized above generalizability of the classifier. To that end, this classifier is trained on the entire annotated dataset, allowing it to be trained on data from each participant, thereby maximizing its effectiveness for the full dataset.

This selected data was divided into continuous, 10 second, non-overlapping segments. The 9 features identified from the feature selection process were computed

for all data segments. Based on these features, the Random Forest classifier provided signal quality ratings for each data segment.

6.2. Quality Assessment Results

The quality ratings for participants were compiled and overall, only 34.8% of the data belonged to class 5 (noise-free). While this is a small proportion of the total data, 34.8% the quality of the signal for that portion is clean enough to be used for cardiovascular analysis. For both non-elderly and elderly participants, a plurality of the data belonged to class 1, representing the most noise corrupted data. This supports the assumption that low signal quality is caused by movement, as the device is located on the wrist, a site which experiences significant motion throughout daily activities. Class 5 data is more prevalent among the elderly subjects, with 38.6% of the data belonging to class 5, compared to only 32.5% for the non-elderly subjects. While mobility was an inclusion criterion for the selection of participants, differences in activity level between the two subject sets were likely responsible for this disparity in data quality. The proportion of class 1 data between the groups is within one percentage point. It was expected that elderly individuals would engage in fewer, less motion-intensive activities throughout the day. Another 15.8% of the PPG data, which belonged to either classes 3 or 4, and could potentially be cleaned sufficiently to derive cardiovascular parameters, thereby enabling 50.6% of the data collected in an ambulatory setting to be useful.

6.3. Quality by Time of Day

To enable a more extensive analysis of the results, the data is subdivided by time of day for both non-elderly and elderly subject sets. The choice of these specific subdivisions were based on a previous work (McCarthy *et al* 2016). A visualization of the signal quality over 24 hours is shown in figure 5. This is an area plot in which the percentage of data belonging to each class, averaged for each 10-minute, non-overlapping time interval, is depicted in a vertically stacked manner. The area occupied by each colour represents the percentage occurrence of its associated class. The plot starts at beginning of the night period, at 22h (10:00pm), and continues for a full 24 hours.

The best data quality was obtained for both subject sets during the night period, from 22h to 7h, which is associated with lower levels of activity as the participants were likely asleep. Also, lower levels of activity were expected at times immediately preceding and succeeding sleep. The markers in figure 5 identify the only time period for which class 5 data consists of more than 50% of the data. This period is seven hours for the non-elderly subject set, and eight hours and ten minutes for the elderly subject set, though it ought to be noted that the elderly subject set has a brief spike in class 1 data at around 7h. The elderly subject set has a longer period during which high quality data is available; however, the results show a higher rate of class 5 data for non-elderly individuals at night. Likely, this is an indication that elderly individuals on average have

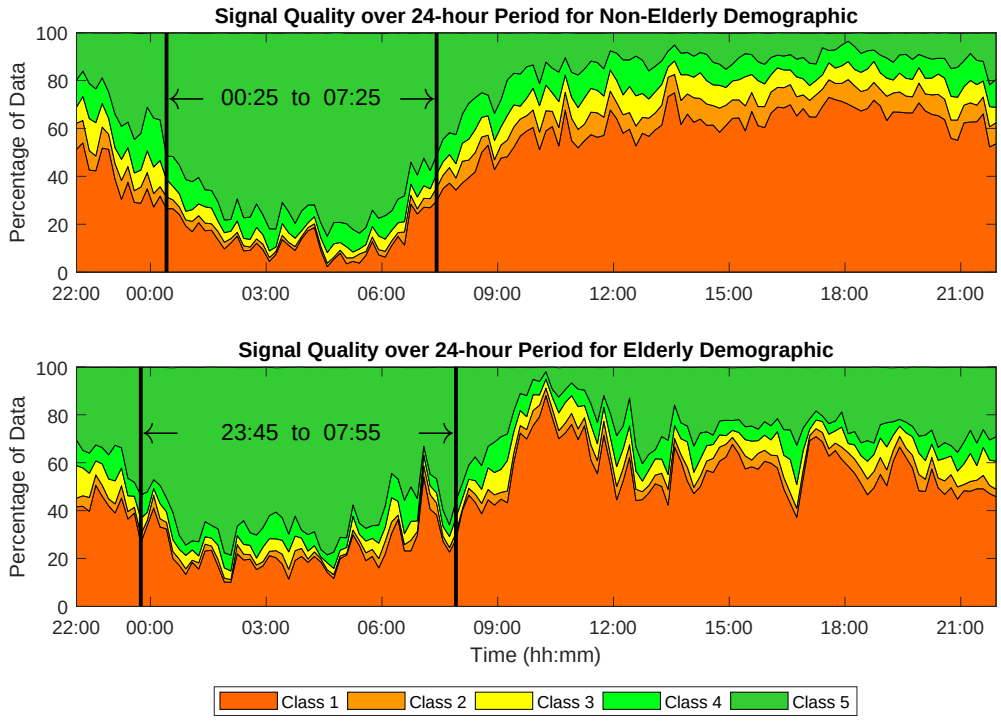

Figure 5. Quality of the PPG over the course of 24 hours for non-elderly and elderly subject sets. The data in the plot is averaged to every ten minutes. The vertical lines delineate the only zones in which class 5 data consists of over 50% of the data.

a longer sleep duration, though may wake up at night more often, thereby resulting in lower rates of class 5 data.

A notable difference between the two subject sets in the portion of class 5 data during the afternoon and evening periods is seen in the visualization in figure 5. The elderly subject set has a class 5 portion 14.9 and 17.3 percentage points higher than the non-elderly category, for the afternoon and evening time periods, respectively. This is likely due to the relatively lower activity levels expected among the elderly subject set. Despite this, only slightly over a quarter of their data during these periods is class 5. Thus, even for elderly individuals, wrist-based PPG technology may not be suitable for continuous monitoring during the day.

7. Discussion

7.1. Limitations

Various limitations were identified in the methodology of the study with regards to the subjects and the data.

The study recruited only mobile subjects, as the scope of the research was to ascertain whether the technology could be used during daily activities. Thus, a segment

of the elderly population which is not mobile, was not included in the study. Higher proportions of class 5 data would be expected in this population segment, for whom it may be a feasible alternative to traditional modalities such as ECG Holter monitors.

Simple class definitions were used to ensure that segments could be easily classified by raters. However, given the subjectivity of visual scoring and the moderate agreement between raters obtained from this study, future studies may opt for stricter class definitions to potentially achieve improvements in rater agreement. Increases in the number of raters may also be trialed in an effort to improve the design of the visual assessment.

The data segments selected for the establishment of the gold standard were chosen randomly from each participant, with 39 ten-second segments being chosen per participant. This random selection resulted in a class imbalance in the data set which was ultimately used in the training and testing of the classifier. Thus, the classifier was trained on more data from classes 1 and 5, compared to the other classes, and performed better at correctly identifying segments of those classes. As a preliminary study, the entire data set was used to maximize the amount of data used to train the classifier. However, future studies could be performed to assess the impact of using a class-balanced training set. This could be achieved by selecting a class-balanced subset from the gold standard data set, and may improve the classifier's accuracy for class 2,3, and 4 data.

The effectiveness of pulse segmentation algorithms used for obtaining the various per pulse statistics was likely inferior in class 1 data segments. Data segments corrupted by motion artifacts would be expected to have a larger number of peaks, which may erroneously be identified as pulses by the pulse segmentation algorithms. However, the lower effectiveness of pulse segmentation for class 1 segments may have been advantageous to the class discriminability, as the per pulse statistics would have yielded different values for these noise peaks. This is demonstrated by the inclusion of these features during feature selection.

Inherent limitations are present in the *Median Relative Power per Pulse* feature, as the relatively low sampling rate of 64 Hz does not allow for high per pulse frequency resolution.

Identification of these methodological limitations form an essential component of this preliminary study, as they can be leveraged for the future development of improved study designs.

7.2. Continuous Ambulatory Monitoring Potential

Signal quality of wrist-based photoplethysmography (PPG) technology was evaluated, as a potential tool for continuous, non-invasive cardiovascular monitoring.

Evaluation of wrist-based PPG technology for continuous, ambulatory use established that the technology provides high quality (class 5) data for only 34.8% of the day, on an average. This in itself may not be sufficient for the detection of

cardiovascular illnesses, however, time analysis of the signal quality reveals that the signal quality improves considerably during periods associated with sleep. Thus, while wrist-based PPG may be unsuitable for continuous monitoring during daily use, it has to the potential to be used in more limited applications, such as night-time monitoring. Despite this limitation, the technology may offer some benefits for early detection if applied for preventative cardiovascular monitoring in relatively lower risk individuals for whom ECG Holter monitoring would not normally be considered due to its cumbersome nature.

Potential improvements may be made to the signal quality through the use of signal processing and hardware adjustments. Future work may seek to identify effective signal processing techniques for the mitigation or elimination of the noise introduced to the signal due to motion artifacts and other sources. Hardware adjustments in the form of LED size, placement, wavelength, light intensity, angle, and the number of LEDs can be studied further for optimal noise reduction. Dynamic feedback systems may also be developed to continuously adjust the LED light intensity or the photodiode sensitivity based to optimize the quality of the obtained signal, potentially even incorporating the accelerometer signal.

Owing to its user-friendliness compared to ECG Holter monitors, wrist-based PPG has potential for long-term ambulatory monitoring. However, the results from this study show that signal quality from the device under test is insufficient in many cases where continuous monitoring is desired.

Acknowledgment

Funding for this research was provided by NSERC Canada and the University of Ottawa Heart Institute.

References

O Abdallah and A Bolz, "Adaptive Filtering by Non-Invasive Vital Signals Monitoring and Diseases Diagnosis," in *Adaptive Filtering Applications*, ch 7, Ed. Lino Garcia Morales, InTech, 2011.

J Allen, "Photoplethysmography and its application in clinical physiological measurement," *Physiol Meas*, 28:R1, 2007.

E Billauer, "Peak detection using MATLAB (non-derivative local extremum, maximum, minimum)," http://www.billauer.co.il/peakdet.html, 2012.

RR Coifman and MV Wickerhauser, "Entropy-based algorithms for best basis selection," *IEEE T Information Theory*, 38:713–718, 1992.

M Elgendi, "Optimal Signal Quality Index for Photoplethysmogram Signals," Bioengineering, 3:21, 2016.

JL Fleiss, "Measuring nominal scale agreement among many raters.," *Psychol Bulletin*, 76:378, 1971.

E Frank, MA Hall, and IH Witten, The WEKA Workbench Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques" 4th Ed, Morgan Kaufmann, 2016.

I Guyon, A Elisseeff, and AM De, "An Introduction to Variable and Feature Selection," *J Machine Learning Research*, 3:1157–1182, 2003.

A Kamal, J Harness, G Irving, and A Mearns, "Skin photoplethysmography a review," *Computer Methods and Programs in Biomedicine* 28:257–269, 1989.

JR Landis and GG Koch, "The measurement of observer agreement for categorical data," *Biometrics*, 33:159–174, 1977.

Q Li and GD Clifford, "Dynamic time warping and machine learning for signal quality assessment of pulsatile signals," *Physiol Meas*, 33:1491–1501, 2012.

H Liang, S Lukkarinen, and I Hartimo, "Heart sound segmentation algorithm based on heart sound envelogram,," pp 105–108 in Proc. Computers in Cardiology, 1997

C McCarthy, N Pradhan, C Redpath, and A Adler, "Validation of the empatica e4 wristband," pp 1–4 in Proc. ISC IEEE EMBS, 2016.

N Pradhan, "Evaluation of the Signal Quality of Wrist-Based Photo- plethysmography," Masters thesis, Carleton University, 2017.

N Pradhan, S Rajan, A Adler, and C Redpath, "Classification of the quality of wristband-based photoplethysmography signals," Proc. IEEE MeMeA, 269–274, 2017.

Y Saeys, I Inza, and P Larraaga, "A review of feature selection techniques in bioinformatics," Bioinformatics, 23:2507–2517, 2007.

JA Sukor, SJ Redmond, and NH Lovell, "Signal quality measures for pulse oximetry through waveform morphology analysis," *Physiol Meas*, 32:369, 2011.

J Wander and D Morris, "A combined segmenting and non-segmenting approach to signal quality estimation for ambulatory photoplethysmography," *Physiol Meas*, 35:2543, 2014.