Skin tone, Confidence, and Data Quality of Heart Rate Sensing in WearOS Smartwatches

Ishita Ray
University of Toronto
Toronto, Canada
ish.ray@mail.utoronto.ca

Daniyal Liaqat
University of Toronto
Toronto, Canada
dliaqat@cs.toronto.edu

Moshe Gabel
University of Toronto
Toronto, Canada
mgabel@cs.toronto.edu

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

***Abstract*—Smartwatches can collect heart rate data unobtrusively and continuously, making them a promising tool for conducting long term studies, monitoring chronic conditions, and providing timely intervention. Healthcare applications, however, require us to understand the reliability of collected readings, both in terms of quality and quantity. The accuracy of optical heart rate (HR) measurements has been studied extensively in recent years, identifying several common causes of errors. For example, previous research has demonstrated that inaccurate HR readings occur more frequently in dark skin as compared to light skin due to melanin absorption. Smartwatches therefore implement a confidence mechanism to estimate reliability of HR readings.**

We study the effect of skin tone on the reliability of confidence estimation of seven consumer-grade WearOS smartwatches. We find that some watches systematically underestimate the reliability of HR readings taken from dark skin, despite no substantial difference in actual error. This results in significantly fewer data points for people with darker skin tones, which can bias downstream applications. We also report a wide variation in how watches implement the same WearOS API for HR collection, with implications for researchers that intend to use them for studies.

***Index Terms*—Health care, Pervasive computing, Sensors**

I. Introduction

Wearable health monitoring technologies have attracted considerable consumer interest over the past few years [1], [2], as they can potentially transform healthcare delivery and research. Accessible, continuous, pervasive, and low-cost health monitoring, these technologies can enable longitudinal health studies, monitoring chronically ill patients, and better healthcare in communities with traditionally poor access to services [3]. In particular, heart rate (HR) monitoring provides important vital signal. For example, HR monitoring is a crucial component of Cardiac Rehabilitation (CR) which improves long-term survival after cardiac events [4].

The adoption of wearable devices for health monitoring creates an urgent need to ensure that their performance is robust for a diverse population [5], [6]. In this paper, we explore the effect of skin tone on the reliability of HR measurements reported by popular smartwatches.

Smartwatches are a promising avenue for continuous HR monitoring: they are programmable, powerful, and unobtrusive sensor platforms that are increasingly popular [7]. In particular, WearOS [8] smartwatches (previously known as Android Wear) are a promising avenue for clinical studies since they are inexpensive, widely available, relatively open, and support third-party applications [9]. Moreover, many WearOS watches integrate HR sensors and provide developers with a standard, portable API for HR sensing.

To measure heart rate, smartwatches use reflective photoplethysmography (PPG): an inexpensive non-invasive optical technique that estimates HR by reflecting light off the skin [10], [11]. The accuracy of wrist-worn PPG HR measurement has been explored extensively in recent years [12]–[15], since clinical monitoring using wearable devices requires confidence in the quality of data. In particular, though early research has shown skin tone can reduce accuracy of PPG HR measurements PPG [11], [16]–[20], recent research shows that the effect is much smaller in modern devices [21], [22], possibly due to better data filtering. Other factors behind erroneous PPG reading include motion [23], [24], change in blood volume [25], and temperature [26]. The WearOS API therefore provides a *confidence* metric: an integer between 0 and 3 that indicates the estimated reliability of individual HR reading; the procedure used to compute this confidence, however, is vendor-defined and often undocumented.

While many works have explored the accuracy of PPG technology for HR sensing, little attention has been devoted to the reliability of the confidence metric, and its interaction with skin tone. Crucially, while the WearOS API is common to all vendors, hardware and software implementations vary greatly. To our knowledge, no prior work has systematically investigated whether confidence reflects actual error across different WearOS watches, and how it is affected by skin tone.

Our contribution: We use seven commercial WearOS smartwatches to collect heart rate readings from 18 participants with a large range of skin tones, and study the quality and quantity of HR measurements and reported confidences.

Our study has two parts: quantitative and qualitative. In the first part, our analysis shows that even when the watch-reported heart rate is in agreement with a gold standard ECG-based device, for dark skin tones many smartwatches report substantially fewer high-confidence data points than for lighter skin tones. This can bias research that relies on data collected by smartwatches since such groups would be under-represented in the dataset and since missing data points can make computing important HR metrics difficult. We further observe that even though the reported confidence of the watches is affected by skin color, across all skin tones, the

TABLE I
SMARTWATCHES IN OUR STUDY.

Watch name	Year	OS	Processor
Moto 360	2014	Android Wear	TI OMAP 3
LG Urbane	2016	Android Wear	Snapdragon 400
Polar 600M	2016	WearOS 2.1	MediaTek MT2601
Huawei Watch 2	2017	WearOS 2.1	Snapdragon Wear 2100
Ticwatch Pro	2018	WearOS 2.1	Snapdragon Wear 2100
Fossil Carlyle HR	2019	WearOS 2.1	Snapdragon Wear 3100
Misfit Vapor X	2019	WearOS 2.1	Snapdragon Wear 3100

highest confidence level represents the data with the lowest actual error (compared to the gold standard) – suggesting that watch-reported confidence can be used to filter out incorrect readings.

In the second part, we find that despite implementing the same data collection API, WearOS smartwatches exhibit widely different and non-obvious behaviors that can affect data quality, and have non-trivial implications for future researchers.

II. STUDY DESIGN

To investigate the effect of skin tone on the reliability of confidence, we used 7 WearOS smartwatches to collect HR data from healthy participants under controlled conditions. The study was approved by the University Health Research Ethics Board (REB No. 38657).

A. Study Equipment

Table I lists watches used in this study. We study seven WearOS-based smartwatches released between 2014 and 2019.

All watches use PPG to generate HR readings and report the heart rate (HR) along with an integer confidence value that can range from 0 to 3. Rather than allowing applications to read HR data on-demand, WearOS calls an application-defined callback function with the reported HR reading and confidence. As we show later, the actual behaviour varies between vendors. For example, some watches report a confidence value of -1 to indicate that the watch is not touching the skin properly. Since we always ensure proper skin contact during data collection, we should expect not to get any data with confidence -1. We explore this further in Section III-D. We use an in-house Android application to collect and transfer recorded data from the smartwatches to a secure server. We collect HR (measured in beats per minute) as well as the watch-reported confidence, augmented with a timestamp with millisecond precision. This data is transferred from the watch to a paired smartphone, and from there transferred to a secure server. At the beginning of each study session, we synchronize clocks by connecting the watches to the smartphone.

The gold standard device used in this study to measure the heart rate is the Zephyr BioHarness 3.0 by Zephyr Technology Corporation [27]. It consists of a strap worn around the chest and an attached module collect heart ECG signals using conductive pads. Heart rate is calculated from ECG data and is reported with confidence values ranging from 0 to 100; we follow a conservative approach and only use readings with confidence of at least 95. The validity of the Zephyr BioHarness in obtaining accurate heart rate data has been confirmed in previous studies [28]. By comparing smartwatch HR readings compared with the HR measured by BioHarness we can correlate watch-reported confidence with actual error. This also helps us monitor our own data collection by discovering outliers (e.g., due to poorly secured watches).

We use a standard quiz based on the Fitzpatrick scale [29]–[31] to group participant skin tones to six groups. As the quiz is somewhat subjective and participants may not be able to provide an exact answer to all questions, we also ask every participant to select a shade card from Pantone skin tone guide [32] that matches their wrist skin color. In cases we could not determine the skin tone group using the incomplete questionnaire, we grouped participants based on the color code chosen from the shade cards.

B. Data Collection procedure

After obtaining informed consent, we collect age, gender, and skin tone using answers to the quiz and the shade card. We then help the participant in wearing he BioHarness band correctly to avoid getting incorrect data due to improper wearing of the chest band. For each participant, data collection is done in two rounds:

1) The participant places both hands on the table. We secure one smartwatch to each wrist, tight enough to ensure proper skin contact while still maintaining comfort.
2) Collect 5 minutes of continuous HR data using both watches. Participants are asked to avoid movement as much as much possible during this time.
3) Remove smartwatches and let participants rest and stretch for 2-3 minutes, before moving to the next pair of watches.
4) Repeat until the participant wore all the watches once on either hand. The order of the watches is random.

After the first round, participants are given a 10 minute break when they can walk around and then rest. We then repeat the procedure for a second round, except smartwatches are now worn on the opposite hand from the first round. This helps reduce confounders such as partipants' dominant hand and differences in circulation between wrists. We then pooled together the data points from both wrists.[1]

Since we aim to study the effect of skin tone on the confidence reported by smartwatches, our study is designed to control other variables known to affect accuracy of wrist-worn PPG-based HR measurements such as light, motion, and elevated heart rate (see Section IV). We collected data at rest and in an upright sitting position. Data collection sessions were carried out in the same room, physical set-up, and artificial lighting. We ensure good skin contact by securing the watches on the participants' wrists using adjustable bands. Finally, we give participants several minutes for their heart rate to settle down before we begin data collection.

[1]We found no significant differences in data from the two wrists.

III. Results

Dependable health monitoring requires a steady stream of reliable data over time. We investigate smartwatch behavior both quantitatively and qualitatively. Our quantitative analysis asks the following research questions:

1) Does watch confidence correlate with the actual error in reported HR reading? Does the highest watch confidence 3 give more accurate data?
2) Can we obtain comparable number of data points across various reported confidence levels and skin tones?
3) Is the correlation between confidence and actual error the same across all skin tones?

Our two main metrics are (a) the average number of HR readings reported by the watch in a fixed time window, and (b) the mean absolute error (MAE) between watch-reported HR and the HR reported by the BioHarness band, measured in beats per minute (BPM). As reported in Section II-A, we discard readings whose BioHarness confidence is below 95.

For qualitative analysis, we investigate the variation between different implementations of the same WearOS APIs.

A. Resulting Dataset

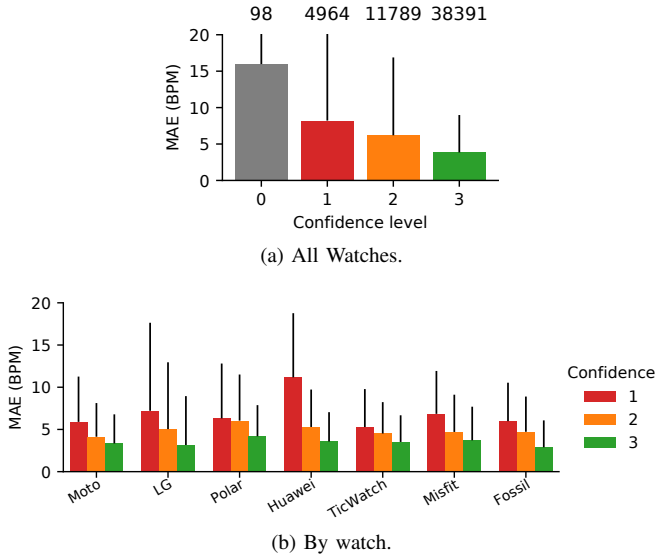
We collected data from 18 healthy participants aged between 18 to 50 years (average age 26.5, 6 male, 12 female); see Section II for collection methodology. Originally we designed the study to include 30 people, 5 people from each of the 6 skin tone groups. Unfortunately, we had to stop data collection midway due to the COVID-19 pandemic. Counting down from the darkest skin tone, we have: two participants with skin tone VI; four with skin tone V; four with skin tone IV; five with skin tone III; two with skin tone II; and one with skin tone I, the lightest skin tone. Data from each participant includes approximately 10 minutes of timestamped HR and confidence values from every smartwatches as well as the BioHarness.

While validating watch data against the data collected from the BioHarness band, we have found that for one participant from skin group IV, the BioHarness band did not collect the heart rate correctly during the entire data collection (its confidence level was below 95 for all data readings). We therefore discard all data from this participant. For skin tone group VI, the TicWatch Pro failed to collect any data from one participant, and for another person from the same group, we lost data for Misfit Vapor X due to watch malfunction.

B. Watch Confidence and Actual Error

We first investigate how well watch-reported confidence correlates with the actual error in HR readings compared to the BioHarness band.

Figure 1(a) shows the mean absolute error (MAE) of heart rate measurement across all the watches and participants for every WearOS confidence level (confidence -1 is not shown, since it means the watch is not touching the skin). Numbers on bars show the number of such data points, and error bars show standard deviations. We observe that confidence 3 gives the most accurate data, and that most of the data points are reported with confidence 3. Ideally, we expect this behavior as



(a) All Watches.

(b) By watch.

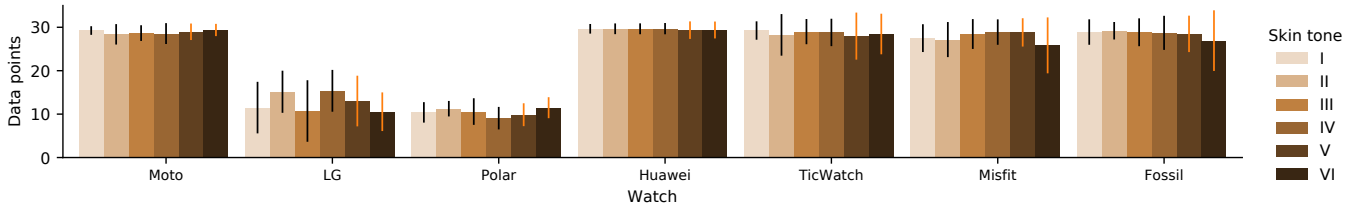
Fig. 1. The MAE for every confidence level. Numbers indicate number of data points. Thin lines shows MAE plus standard deviation.

our data collection procedure (Section II-B) is designed to reduce interference from movement, elevated heart rate, ambient light, and improper skin contact. Since confidence level 0 has very few data points (less than 0.2% of the whole dataset), we exclude it from the rest of our analysis. For confidence levels $1-3$ MAEs are consistent with the MAE reported at rest by Bent et al. [21] (our MAE is slightly lower since our study protocol includes plenty of rest before measurement resulting in lower HR). Hence the WearOS smartwatches used in our study provide comparable performance to popular non-WearOS smartwatches. Figure 1(b) shows the MAE of each watch individually. We again observe that for every watch, data reported with confidence 3 has the lowest error. **We conclude that watch-reported confidence correlates with the actual error, and that HR reported with confidence 3 is the most accurate across all watches.**

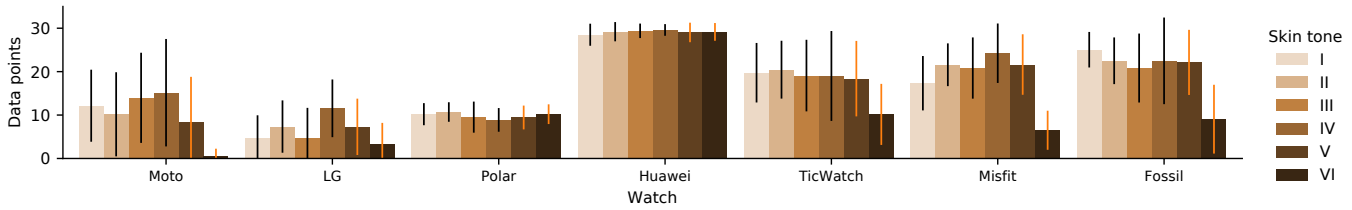
C. Effect of Skin Tone on Confidence, Quality, and Quantity

Figure 2(a) shows the average number of data points (readings) reported by each watch in every 30-second window for different skin tones. Superficially, skin tone seems to have little effect on the number of data points reported by each watch.

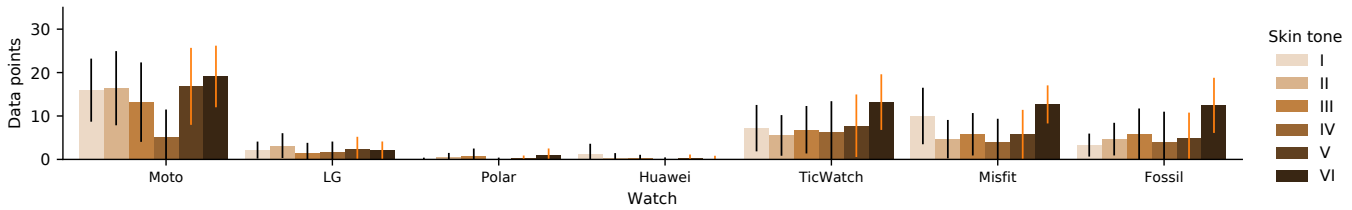
Most researchers and applications, however, require reliable, high-confidence data. Having observed that confidence level 3 gives the most accurate data for all the watches, we look into the number of such data points reported in a fixed time window for these watches. The number of data points reported with confidence 3 (Figure 2(b)) is substantially smaller for skin tone VI than for the lighter skin tones. For the Moto, Ticwatch, Fossil, and Misfit watches the number of reliable (confidence 3) HR readings is smaller by almost 50% or more for skin tone VI than for lighter skin tones. The LG watch exhibits a less substantial drop, while the number of high-confidence data points for Polar and Huawei watches is unaffected by skin tone. We see the opposite trend rise for low confidence



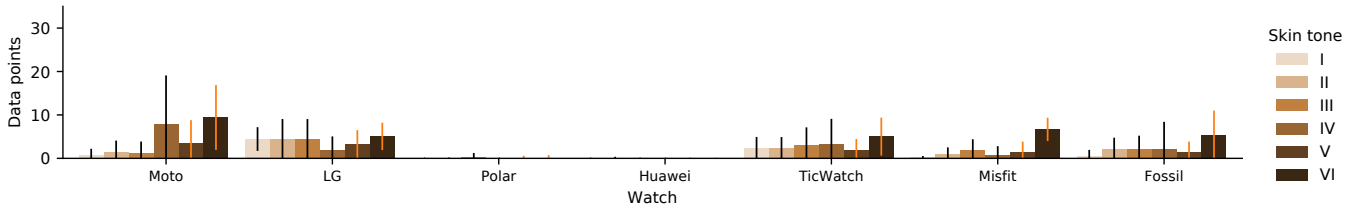
(a) All confidences.



(b) Only confidence 3.



(c) Only confidence 2.



(d) Only confidence 1.

Fig. 2. Average number of data points reported every 30 seconds by every watch with all confidence levels (a), with confidence 3 only (b), confidence 2 only (c), and confidence 1 only (d). Error bars show standard deviation.

levels, where that darker skin tones have more readings per 30 second window (Figures 2(c) and 2(d)). **In conclusion, four out of the seven watches report significantly fewer HR data with confidence** 3 **for darker skin tones, even though skin tone has little impact on the total number of reported data by the watches.**

Given that confidence 3 is associated with the most accurate data, these findings raise the possibility that the drop in the number of high confidence data points may be caused by less accurate HR readings from darker skin tones. Figure 3 shows the MAE of HR measurements across all watches for all confidence levels for different skin tones (top), and for high confidence data only (bottom). As expected, using only readings with confidence 3 have lowered the actual error for all watches. However, this improvement in accuracy is equally spread across all skin tones for every watch. **Hence, the issue is not the lack of high-quality data, but that the confidence estimation in some watches is poor for dark skin tones.**

To further support this conclusion, we explore the distribution of confidences for accurate data points, those with low MAE. Since Figure 1(a) shows that mean absolute error across all skin tones and watches is approximately 4 BPM, and so we conservatively consider as accurate only HR readings with absolute error below 2 BPM compared to the BioHarness (3% of nominal HR of 70 BPM). Ideally, most accurate data points would be reported with confidence level 3. Unfortunately, we observe for darker skin tones, this is not the case. Figure 4 shows the percentage of accurate data points for each confidence in every skin tone group. It includes data from the four watches with substantial drops (above 1 standard deviation) in the number of data points for darker skin: Moto, Ticwatch, Fossil, and Misfit. While most readings have confidence 3 for skin tones I to V, for skin tone VI the majority of data has confidence 2, despite only including accurate readings. **We conclude that confidence is miss-calibrated for dark skin tones: most WearOS watches are too conservatively. This results in fewer high-confidence readings from darker skin tones, despite no difference in true accuracy.**

Note that allowing low confidence readings for people with darker skin tones may not be a practical solution for data

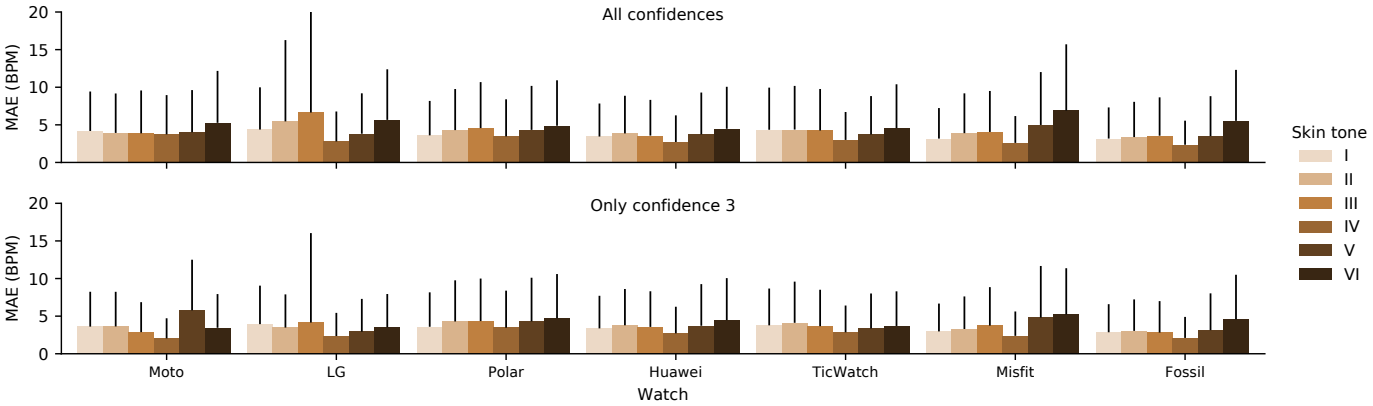


Fig. 3. MAE for all the watch reported data (top) and only high-confidence data (bottom). Lines show standard deviation.

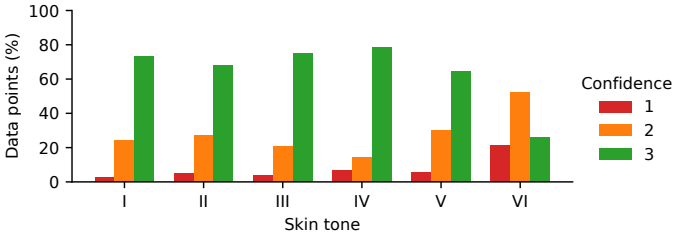


Fig. 4. The distribution of confidence for data points with absolute error ≤ 2 from the Moto, TicWatch, Fossil, and Misfit watches.

TABLE II
DIFFERENCES IN API IMPLEMENTATION ACROSS WEAROS WATCHES.

	Moto	**LG**	**Polar**	**Huawei**	**TicWatch**	**Misfit**	**Fossil**
reported confidence	0–3	0–3	0–3 (rarely -1)	-1–3	0–3	0–3	0–3
phantom data	Yes	Yes	Yes	No	No	No	No
confidence fluctuation	Yes	Yes	No	No	Some	Some	Some
data reporting frequency	1 Hz	variable	0.5 Hz	1 Hz	1 Hz	1 Hz	1 Hz
suspends data collection app	No	No	No	No	Yes	Yes	Yes

collection. We conducted our study in a controlled environment that minimizes all other potential sources of error such as movement and ambient light. In real-world settings, low confidence readings are likely to be inaccurate even for the darker skin tones. We intend to explore this in future work.

D. Inconsistent API Implementation

Beyond the quantitative effect of skin tone on the watch confidence, another factor that can make the utilization of watch reported heart rate challenging are qualitative differences in API implementation across watches. Developing smartwatch applications for health monitoring or HR data collection means following the documentation of WearOS APIs. However, it turns out that different watches implement the same API differently, and researchers should be aware of those differences while collecting or using WearOS HR data.

Table II summarizes differences in how different WearOS watches implement the HR collection API. Differences likely to impact data quality are highlighted in red. Differences that may require researcher or developer attention are highlighted in yellow. Developers of WearOS smartwatch applications that use or collect heart rate data should keep these in mind.

Reported Confidence and Phantom Data: While WearOS specifies a confidence scale, it does not provide a clear criteria for each confidence level, and every vendor uses a different method to decide what each level means and how to determine it. WearOS API also allows -1 confidence, intended to report loss of contact with skin. Though this potentially allows applications to determine whether the watch is worn, in practice not all watches implement it. We tested how the seven watches behave in absence of skin contact. The Ticwatch, Fossil, and Misfit watches never report -1 confidence explicitly, but instead, they do not report any HR data if the watch is lacking skin contact This suggests that these watches can identify whether a watch is worn or not, but do not report it to the application. On the other hand, we found the Huawei watch to consistently report confidence level -1 in the absence of skin contact. While the Polar watch occasionally does report -1 confidence, we have found that this watch can report phantom HR data with confidence 3 for up to ten seconds even when no one is wearing it. The Moto and LG watches never report -1 confidence and appear to be unable to detect if the watch is worn. As with the Polar watch, we have observed that the Moto watch reporting phantom HR data with watch confidence 3 for several seconds even when no one is wearing the watch. The LG watch also occasionally reports the highest confidence while it lacks skin contact. **Hence, despite the API documentation, researchers should not assume that watches can reliably identify and notify when the watch is being worn. Moreover, for some watches occasionally report phantom HR readings with high confidence. Depending on the use-case, researchers may need to develop their own mechanisms to identify whether the watch is worn and data is reliable.**

Confidence Fluctuation: Given our data was collected in a controlled environment, we expected the confidence of continuous HR readings to stabilize after a short period. But in practice, only the Polar and Huawei watches provided stable confidence of 3 over time. The confidence from Moto and LG watches fluctuated regularly between 2 and 3. On Fossil, Misfit, and TicWatch confidence dropped from 3 to either 2 or 1 a few times every minute and typically took $8-10$ seconds to return to 3. We have also observed that at the beginning of each data collection session, some watches report a few data points with confidence 0. Among them, only Polar and Moto watch report HR as 0 along with confidence 0 for $3-5$ seconds. **Researchers should not assume confidence fluctuations indicate changes in the environment, movement, or activity.**

Data Reporting Frequency: Figure 2 shows most watches report one reading every second (1 Hz). The Polar watch reports data once every two seconds, while LG shows high jitter in HR reporting frequency – the times between successive reports fluctuate widely, with an average of one reading every two seconds. Such behavior can introduce bias to analysis where a decision made from the data is also dependent on the amount of data along with the quality. **Researchers should be prepared to accept measurements taken at different frequencies that can require aligning the data before doing any analysis. Moreover, they should be prepared to deal with or at least detect irregularly reported data from watches like the LG, for example by leveraging timestamp reported by theWearOS API.**

Data collection App suspension: We found that continuous HR logging in a third party application can be challenging in some watches due to aggressive power management. The TicWatch and Fossil watches consistently shut down our data collection app (Section II-A) 2 minutes into the run even when the watches are not on any energy-saving mode. The Misfit watch also shows this behavior infrequently. We had to manually keep the watch screen alive to prevent this. **Researchers should take measures to detect and, if possible, prevent such issues. Careful testing requires a veriety of devices, since suspend behavior of one watch does not necessarily match that of other watches.**

IV. Related Work

Fallow et al. [16] investigate the interaction between light wavelength and skin tone at rest and during exercise, and demonstrate that devices are better at detecting pulses using green light at rest, and green or blue light during exercises. In later work, Spierer et al. [17] present a validation study of the Omron HR500U and Mio Alpha wrist-worn dedicated heart rate monitors across a range of physical tasks. Along with physical movement, the authors also considered the effect of photosensitivity of skin on the correctness of HR readings and had reported that error rate increased linearly with less photosensitive (darker) skin for the Mio Alpha, but not for HR500U. However, this study did not include people from all skin tone groups. Recently, Bent et al. [21] extensively explored the effect of skin type on the error in heart rate measurement, both at rest and in motion, in consumer-grade wrist-worn devices such such as the Apple Watch 4, Fitbit Charge 2, and the Garmin Vivosmart 3. Focusing on exercises, they found that motion has a higher effect on HR measurement error than skin tone. Our mean absolute errors are consistent with their findings at rest. Horton et al. [22] validated the WearOS-based Polar M600 watch against conventional ECG for different activities such as at rest, during various physical activities, and during recovery. They observed a tendency to underestimate HR during intense activity overestimation it when intensity decreases, but no statistically significant interaction with gender, body mass index, skin type, or wrist size. As with Spierer et al. [17] this study did not have participants from all the skin tone groups. Other works [24], [33], [34] similarly explore the accuracy of smartwatch HR readings compared to a gold standard device.

Unlike prior works, which focus on HR measurement error, we focus on the WearOS reliability reporting mechanism (i.e., confidence). We systematically study the effect of skin tones under carefully controlled conditions, and point out how mechanisms that increase accuracy can cause other data quality issues. Moreover, we specifically focus on WearOS, the most popular smartwatch OS among vendors, and investigate variation in API implementation across watches. To our best knowledge, our work is the first to investigate those questions.

V. Conclusion

Smartwatches are a promising technology for pervasive, low-cost heart rate (HR) monitoring. While their measurement accuracy has been extensively studied, other data quality issues such as data quantity and reliability remain under-explored. We systematically study the reliability of watch-reported confidence values of seven WearOS smartwatches from 18 participants with six different skin tones groups. We find that for several watchers, confidence is poorly calibrated for dark skin tones and does not reflect the true accuracy of reported HR readings. This can result in under-representing people with darker skins in data collected by such smartwatches, which in turn can bias downstream research (e.g., machine learning models). Finally, we find substantial variation in WearOS API implementations, which can invisibly impact data quality.

The ongoing COVID-19 pandemic has forced us to stop the study early. It should therefore be considered exploratory and our results preliminary. We purposefully avoid statistical tests given its low population. Nevertheless, we believe our early findings are valuable to researches and developers that aim use smartwatches for pervasive HR monitoring. We intend to expand our study to cover more people, as well as systematically explore other factors that affect watch confidence such as movement, light, and temperature. Future work could also focus on using machine learning to replace or augment watch confidence using other sensors such as IMU and light sensors [35], [36].

REFERENCES

[1] S. R. Islam, D. Kwak, M. H. Kabir, M. Hossain, and K.-S. Kwak, "The internet of things for health care: a comprehensive survey," *IEEE Access*, vol. 3, pp. 678–708, 2015.

[2] R. Rawassizadeh, B. A. Price, and M. Petre, "Wearables: Has the age of smartwatches finally arrived?" *CACM*, vol. 58, no. 1, pp. 45–47, 2014.

[3] M. Chan, D. Estève, J.-Y. Fourniols, C. Escriba, and E. Campo, "Smart wearable systems: Current status and future challenges," *Artif. Intell. Med.*, vol. 56, no. 3, pp. 137–156, 2012.

[4] S. M. Dunlay, Q. R. Pack, R. J. Thomas, J. M. Killian, and V. L. Roger, "Participation in cardiac rehabilitation, readmissions, and death after acute myocardial infarction," *Am. J. Med.*, vol. 127, no. 6, pp. 538–546, 2014.

[5] P. J. Colvonen, P. N. DeYoung, N.-O. A. Bosompra, and R. L. Owens, "Limiting racial disparities and bias for wearable devices in health science research," *Sleep*, vol. 43, no. 10, 2020.

[6] S. Harrison, "How accurate is your commercial fitness tracker?" https://themarkup.org/ask-the-markup/2020/05/21/how-accurate-is-your-commercial-fitness-tracker, 2020.

[7] A. M. Research, "Global opportunity analysis and industry forecast, 2020-2027," "https://www.alliedmarketresearch.com/smartwatch-market", April 2020.

[8] Google, "WearOS," "https://en.wikipedia.org/wiki/Wear_OS".

[9] R. Liu, L. Jiang, N. Jiang, and F. X. Lin, "Anatomizing system activities on interactive wearable devices," in *Proc. APSys*, 2015, pp. 1–7.

[10] C. Wang, Z. Li, and X. Wei, "Monitoring heart and respiratory rates at radial artery based on PPG," *Optik*, vol. 124, no. 19, pp. 3954–3956, 2013.

[11] M. Nitzan, A. Romem, and R. Koppel, "Pulse oximetry: fundamentals and technology update," *Med. Devices (Auckl.)*, vol. 7, p. 231, 2014.

[12] R. Hailu, "Fitbits, other wearables may not accurately track heart rates in people of color," "https://www.statnews.com/2019/07/24/fitbit-accuracy-dark-skin/", 2019.

[13] E. Jo, K. Lewis, D. Directo, M. J. Kim, and B. A. Dolezal, "Validation of biofeedback wearables for photoplethysmographic heart rate tracking," *J. Sports Sci. Med.*, vol. 15, no. 3, p. 540, 2016.

[14] R. K. Reddy, R. Pooni, D. P. Zaharieva, B. Senf, J. El Youssef, E. Dassau, F. J. Doyle III, M. A. Clements, M. R. Rickels, S. R. Patton *et al.*, "Accuracy of wrist-worn activity monitors during common daily physical activities and types of structured exercise: evaluation study," *JMIR mHealth and uHealth*, vol. 6, no. 12, p. e10338, 2018.

[15] F. Sartor, J. Gelissen, R. Van Dinther, D. Roovers, G. B. Papini, and G. Coppola, "Wrist-worn optical and chest strap heart rate comparison in a heterogeneous sample of healthy individuals and in coronary artery disease patients," *BMC Sports Sci. Med. Rehabil.*, vol. 10, no. 1, 2018.

[16] B. A. Fallow, T. Tarumi, and H. Tanaka, "Influence of skin type and wavelength on light wave reflectance," *J. Clin. Monit. Comput.*, vol. 27, no. 3, pp. 313–317, 2013.

[17] D. K. Spierer, Z. Rosen, L. L. Litman, and K. Fujii, "Validation of photoplethysmography as a method to detect heart rate during rest and exercise," *J. Med. Eng. Technol.*, vol. 39, no. 5, pp. 264–271, 2015.

[18] L. Yan, S. Hu, A. Alzahrani, S. Alharbi, and P. Blanos, "A multi-wavelength opto-electronic patch sensor to effectively detect physiological changes against human skin types," *Biosensors*, vol. 7, no. 2, p. 22, 2017.

[19] J. Spigulis, L. Gailite, A. Lihachev, and R. Erts, "Simultaneous recording of skin blood pulsations at different vascular depths by multiwavelength photoplethysmography," *Applied optics*, vol. 46, no. 10, pp. 1754–1759, 2007.

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

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

[22] J. F. Horton, P. Stergiou, T. S. Fung, and L. Katz, "Comparison of Polar M600 optical heart rate and ECG heart rate during exercise," *Med. Sci. Sports Exerc.*, vol. 49, no. 12, pp. 2600–2607, 2017.

[23] A.-M. Tăuţan, A. Young, E. Wentink, and F. Wieringa, "Characterization and reduction of motion artifacts in photoplethysmographic signals from a wrist-worn device," in *EMBC*, 2015, pp. 6146–6149.

[24] J. Xie, D. Wen, L. Liang, Y. Jia, L. Gao, and J. Lei, "Evaluating the validity of current mainstream wearable devices in fitness tracking under various physical activities: comparative study," *JMIR mHealth and uHealth*, vol. 6, no. 4, p. e94, 2018.

[25] T. Tamura, Y. Maeda, M. Sekine, and M. Yoshida, "Wearable photoplethysmographic sensors—past and present," *Electronics*, vol. 3, no. 2, pp. 282–302, 2014.

[26] I. C. Jeong, H. Yoon, H. Kang, and H. Yeom, "Effects of skin surface temperature on photoplethysmograph," *J. Healthc. Eng.*, vol. 5, 2014.

[27] WearableTech, "Zephyr bioharness 3.0," "https://wearabletech.io/zephyr-bioharness-3/", 2019.

[28] J. A. Johnstone, P. A. Ford, G. Hughes, T. Watson, and A. T. Garrett, "BioHarness™ multivariable monitoring device: part. I: validity," *J. Sports Sci. Med.*, vol. 11, no. 3, p. 400, 2012.

[29] T. B. Fitzpatrick, "The validity and practicality of sun-reactive skin types I through VI," *Arch. Dermatol.*, vol. 124, no. 6, pp. 869–871, 1988.

[30] S. Sachdeva *et al.*, "Fitzpatrick skin typing: applications in dermatology," *Indian J. Dermatol. Venereol. Leprol.*, vol. 75, no. 1, p. 93, 2009.

[31] D. Creation, "Fitzpatrick skin type quiz." "https://www.devotedcreations.com/docs/The_Fitzpatrick_Skin.pdf", 2017.

[32] Pantone, "Pantone skin-tone guide," "https://www.pantone.com/skintone-guide/".

[33] J. Pietilä, S. Mehrang, J. Tolonen, E. Helander, H. Jimison, M. Pavel, and I. Korhonen, "Evaluation of the accuracy and reliability for photoplethysmography based heart rate and beat-to-beat detection during daily activities," in *EMBEC & NBC 2017*, 2018, pp. 145–148.

[34] S. E. Stahl, H.-S. An, D. M. Dinkel, J. M. Noble, and J.-M. Lee, "How accurate are the wrist-based heart rate monitors during walking and running activities? are they accurate enough?" *BMJ Open Sport Exerc. Med.*, vol. 2, no. 1, 2016.

[35] C. Phillips, D. Liaqat, M. Gabel, and E. de Lara, "WristO_2: Reliable peripheral oxygen saturation readings from wrist-worn pulse oximeters," in *PerCom Workshops*, 2021.

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