

Continuous Stress Detection Using the Sensors of Commercial Smartwatch

Pekka Siirtola

pekka.siirtola@oulu.fi

Biomimetics and Intelligent Systems Group
University of Oulu, Finland

ABSTRACT

Stress detection is becoming a popular field in machine learning and this study focuses on recognizing stress using the sensors of commercially available smartwatches. In most of the previous studies, stress detection is based on partly or fully on electrodermal activity sensor (EDA). However, if the final aim of the study is to build a smartwatch application, using EDA signal is problematic as the smartwatches currently in the market do not include sensor to measure EDA signal. Therefore, this study surveys what sensors the smartwatches currently in the market include, and which of them 3rd party developers have access to. Moreover, it is studied how accurately stress can be detected user-independently using different sensor combinations. In addition, it is studied how detection rates vary between study subjects and what kind of effect window size has to the recognition rates. All of the experiments are based on publicly available WESAD dataset. The results show that, indeed, EDA signal is not necessary when detecting stress user-independently, and therefore, commercial smartwatches can be used for recognizing stress when the used window length is big enough. However, it is also noted that recognition rate varies a lot between the study subjects.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; • **Applied computing** → **Life and medical sciences**.

KEYWORDS

Wearable sensors, biosignals, stress detection, machine learning

ACM Reference Format:

Pekka Siirtola. 2019. Continuous Stress Detection Using the Sensors of Commercial Smartwatch. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and the 2019 International Symposium on Wearable Computers (UbiComp/ISWC '19 Adjunct)*, September 9–13, 2019, London, United Kingdom. ACM, New York, NY, USA, 4 pages. <https://doi.org/10.1145/3341162.3344831>

1 INTRODUCTION

There is evidence that well-being at work and working efficiency are connected together [1]. This is because, employees that are feeling well at work have better engagement, more motivated and they have less sick days. One of the factors leading to reduced work well-being and working efficiency is stress. To increase productivity, it should be studied what causes stress at work, and before this can be done, methods to measure stress needs to be studied. In this article, it is studied how to measure stress using the sensors of commercial smartwatches.

Stress detection using wrist-worn sensors has recently been studied widely, and it has been shown in the previous studies that stress can be recognized with quite reliably. For instance, in [2] stress at laboratory conditions was detected with the accuracy of 83% using skin temperature, electrodermal activity, heart rate, blood volume pulse and accelerometer signals. In fact, most of the studies use, among other bio-sensors, electrodermal activity (EDA, a.k.a. galvanic skin response) signal in the recognition process [3]. The main problem of EDA is that if the aim is to build a stress detection application for off-the-shelf smartwatches, the recognition cannot be based on EDA as the smartwatches currently at the market do not include sensor to measure EDA.

There are some studies where EDA is not used in the stress detection process, and instead of EDA, other bio-signals have been used. For instance, in [4] only accelerometer and optical heart rate sensor data were used for unsupervised stress detection. However, there is no systematic study on how accurately stress can be detected user-independently using different sensor combinations. Some of the work was already done in [5] where some of the sensor combinations were compared. In fact, this article extends [5] by giving more insights to publicly open WESAD data set and stress detection

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

UbiComp/ISWC '19 Adjunct, September 9–13, 2019, London, United Kingdom

© 2019 Association for Computing Machinery.

ACM ISBN 978-1-4503-6869-8/19/09...\$15.00

<https://doi.org/10.1145/3341162.3344831>

Table 1: Smartwatch markets are dominated by three operating systems. The SDKs of these support different sensors.

Sensor / OS	Tizen	WearOS	watchOS	E4
ACC	X	X	X	X
ST	X	X	X	X
EDA	-	-	-	X
BVP	X	-	-	X
HR	X	X	X	X
HRV	X	-	-	X

using wearable sensors. Especially, this article concentrates on studying how well stress can be detected using the sensors included to commercial smartwatches. Moreover, it is studied how the recognition accuracy varies between study subjects, as well as, how the window size effects to the recognition rates.

2 EXPERIMENTAL DATASET

This study is based on publicly open data set called WESAD [5]. It contains data from 15 participants measured using Empatica E4 wrist-worn device [6]. This device includes accelerometers (ACC), as well as, sensors to measure skin temperature (ST), electrodermal activity (EDA), blood volume pulse (BVP), heart rate (HR), and heart rate variability (HRV). In addition to E4 data, WESAD includes data from chest-worn RespiBAN device and participants also filled in questionnaires related to their feelings during the data gathering session. However, in this study, only E4 data is used as the focus of this article is in wrist-worn sensors.

In the data gathering, data from three different affective states (stress, amusement, and relaxed) were collected. Approximately the length of the stressed situation was 10 minutes, amused situation 6.5 minutes and, relaxed situation was 20 minutes. In this study, amusement and relaxed states were combined as one. Therefore, the studied problem was binary (stressed vs. non-stressed). More detailed information about the dataset can be found from [5].

3 SMARTWATCH SENSORS

Currently, the smartwatch market is dominated by three big operating systems: Tizen by Samsung, WearOS by Google, and watchOS by Apple. All of these operating systems have own strength and weaknesses, but in this study the main interest is which sensors they include and which one's third party application developers have access to. In Table 1, these operating systems are compared to Empatica E4, which is sensor-wise the most comprehensive wrist-sensor available. However, Empatica is not a smartwatch, instead it is a device

that is designed to research usage, and therefore, it is expensive and it is not a consumer product that can be bought from store.

The comparison shows that from the commercial smartwatches, Tizen is superior to WearOS and watchOS sensor-wise. In fact, only Tizen provides access to heart rate variability which is, according to several studies, the most important information source when detecting different symptoms and diseases. Moreover, only Empatica E4 includes a sensor to measure electrodermal activity (EDA), which is in many studies used to detect stress. Therefore, if the aim is to detect stress user-independently using commercial smartwatches, it needs to be studied if it is possible to do that without EDA sensor, and if EDA sensor is not obligatory, which sensors are?

4 METHODS

In this article, the detection of stress is based on a recognition process traditionally used in human activity recognition studies. This process contains three phases: data gathering, training and detection [7]. The data gathering process is passed by using open data set. For the model training signals were divided into windows, and from these windows, features were extracted. To study the effect of window size, different window sizes (15s, 30s, 60s, 90s, and 120s) were compared. However, in each case, the length of slide was the same, 0.25 seconds. In the original WESAD article [5] this same slide was used with window size of 60 seconds. Moreover, the features extracted from the windows were the same as the ones used in [5]. However, in the original article, features were extracted from EDA, BVP, temperature and accelerometer signals, but not from heart rate signal as it was done on this article.

Different window sizes and features are tested using 3 different classifiers: LDA (linear discriminant analysis), QDA (quadratic discriminant analysis), and RF (Random Forest). These classifiers were selected as they have shown good performance when detecting human activities based on wearable sensor data.

5 EXPERIMENTS

For the experiments, the data was divided into training and testing sets using leave-one-out method as the purpose is to build user-independent models to detect stressed and non-stressed stages. This means that one person's data in turn was used for testing and other 14 for training. Feature selection was not used and the results in the tables present balanced accuracies, which means that accuracy for both classes was calculated separately and balanced accuracy is an average of these.

Table 2: Window size has a noticeable effect on the recognition rates. Results using LDA classifier and the data from all available biosensors.

Window size	Balanced accuracy	Standard deviation
15s	81.2	13.1
30s	79.9	13.0
60s	81.7	13.3
90s	83.2	11.6
120s	84.0	12.9

Table 3: Average recognition results accuracies (standard deviation in parentheses) using different classifiers and sensor combinations.

Sensors	LDA	QDA	RF
ACC+EDA+ST+BVP+HR	83.4 (11.3)	76.8 (11.5)	80.5 (16.4)
EDA+ST+BVP+HR	84.0 (12.9)	72.9 (15.6)	80.7 (14.4)
ST+BVP+HR	87.4 (10.4)	84.9 (14.3)	82.4 (15.7)
EDA+BVP+HR	82.2 (13.1)	78.4 (12.9)	80.8 (12.8)
EDA+ST+HR	84.4 (12.4)	72.1 (15.9)	81.3 (13.8)
EDA+ST+BVP	78.8 (15.2)	81.6 (15.7)	81.0 (16.6)
ST+HR	82.5 (12.4)	81.1 (12.6)	77.5 (16.9)
EDA+BVP	77.8 (13.7)	71.7 (15.4)	79.5 (14.5)
BVP+ST	82.9 (11.5)	83.3 (15.0)	79.8 (18.4)
EDA+ST	78.6 (15.4)	72.1 (15.9)	79.7 (12.3)
EDA+HR	82.7 (14.3)	74.7 (13.1)	78.8 (13.3)
BVP+HR	83.8 (11.8)	78.7 (14.9)	82.5 (10.5)
EDA	73.5 (16.6)	69.7 (15.7)	78.3 (13.6)
BVP	81.2 (11.4)	67.9 (15.2)	81.4 (9.2)
ST	75.2 (17.5)	68.3 (18.1)	66.9 (18.0)
HR	77.0 (18.3)	73.3 (16.0)	73.2 (13.6)
ACC	71.0 (8.9)	69.2 (10.8)	80.0 (9.2)

The effect of window size

The results presented in Table 2 show the effect of window size to the detection rate. In this case, LDA and all the features biosignals (= features from skin temperature, electrodermal activity, heart rate, and blood volume pulse) were used in the classification process. Tested window sizes varied from 15 to 120 seconds, longer window sizes were not experimented as the size of the data set became a limiting factor. Based on Table 2, window size has effect to the detection accuracies and long window size provides better results than short ones. Interestingly, though the methods used to recognize stress are quite similar to ones used to detect human activities, stress detection requires longer window than human activities, which are normally detected using window size of 2-10 seconds. This shows the difference of motion and biosignal analysis, the type of motion can change rapidly while affective states do not change that fast.

According to the Table 2, 120 seconds window provides the best results. Therefore, this window size is used in the following experiments when different classifiers and sensor combinations are compared.

Comparing different sensor combinations and classifiers

Table 3 compares the recognition rates using three different classifiers and all the possible biosensor combinations. Firstly, it can be noted that in the most of the cases LDA performs slightly better than Random Forest and much better than QDA. In fact, when it was noted that LDA performs better than Random Forest, it was decided that there is no need to run experiments several times using Random Forest, though, depending on the run there is slight variation in the results provided by Random Forest. On the other hand, the results of LDA and QDA are not depended on the run. Secondly, it can be noted that with each classifier, the best recognition rate (LDA: 87.4%, QDA: 84.9%, RF: 82.4%) was obtained using a combination of skin temperature, BVP, and heart rate signals. Therefore, when stress is detected, it is not necessary to use EDA signal in the recognition process. In fact, according to the experiment, better detection rates can be obtained when it is not used at all. Moreover, the importance of heart-based biosignals (BVP and heart rate) can be noted from the results, for instance by using only BVP signal, over 80% recognition rate can be obtained. Additionally, it is worth noticing that accelerometer data is not needed in stress detection, better results can be obtained when biosensor data is not supported by accelerometer data when using LDA or Random Forest as the classifier. This shows that while accelerometer provides information about motion, it is not that informative source of data when it comes to detecting affective states.

Table 4 shows how the recognition rate varies between the study subjects. There is a lot of variation, the lowest accuracy is 62.6% and highest 98.8%. In fact, to reduce the variation, it should be studied if the questionnaires explains why the stress reactions of study subjects 14 & 15 are more difficult to recognize than the one's of other subjects.

6 DISCUSSION

The results presented in Table 2 show the effect of window size to the recognition rates. In overall, it can be noted that by using long windows, better recognition rates can be obtained than using short ones. In this case, window of length 120 seconds provided the best results, and it was used in the following experiments. It should be further experimented if the recognition of other affective states, such as amusement, also requires long windows. Moreover, longer dataset would allow experimenting with window sizes longer than

Table 4: The best recognition rates were obtained using LDA classifier and features from skin temperature, BVP and heart rate signals. However, the recognition rate varies a lot between the study subjects.

Study subject	Accuracy
2	84.5
3	85.7
4	96.2
5	87.7
6	93.8
7	86.6
8	91.7
9	80.0
10	93.5
11	91.2
13	98.8
14	67.7
15	62.6
16	93.3
17	97.4

120 seconds. Now the size of the data set did not allow experimenting with longer window sizes. For instance, in [2] window size of 20 minutes was used.

The main aim of the article was to study if it is possible to detect stress using the sensors of commercial smartwatches which do not include EDA sensor. The answer is yes. Tizen SDK provides access to skin temperature, BVP, and heart rate, and according to Table 3 this sensor combination provides recognition accuracy of 87.4% when LDA classifier is used. In fact, recognition rate slightly drops if EDA signal is used along with these these signals. The SDKs of WearOS and watchOS are more limited sensor-wise, the only provide access to body temperature and heart rate sensors. With this combination, stress can be detected with accuracy of 82.5% when using LDA classifier. Moreover, LDA classifier seems to provide better results than QDA and Random Forest.

The main problem in stress detection that recognition rate varies a lot between the study subjects, see Table 4. Therefore, future work includes studying if the variation could be reduced by using personalized recognition models (for instance using method presented in [8]) instead of user-independent or if the reason for variation can be found from the questionnaires. Moreover, the author believes that results can be improved by using feature selection.

7 CONCLUSION

Continuous stress detection using sensors of a smartwatch was studied in this article. Firstly, the SDKs of the different smartwatch operation systems were studied to find out what

data sources are available to third party developers. Tizen SDK was found the most comprehensive, it provides access to motion, ST, BVP, and HR sensors. Secondly, it was studied how the window size effects to stress detection rates. It was noted that long window (120s) provide more accurate results than shorter one's. Thirdly, the main aim was to study if it is possible to detect stress user-independently using the sensors of commercial smartwatches which do not include EDA sensor. Leave-one-out method was used in the classification process and it was actually noted that more accurate recognition models can be trained if EDA is not used at all. The best results (balanced accuracy 87.4%) were obtained using a combination of LDA classifier and ST, BVP, and HR sensors. However, while the average recognition rate is high, the recognition rate varies a lot between the study subjects.

ACKNOWLEDGEMENT

This research is supported by the Business Finland funding for Reboot IoT Factory-project (www.rebootiotfactory.fi). Author is also thankful for Infotech Oulu.

REFERENCES

- [1] N. Renee Baptiste, "Tightening the link between employee wellbeing at work and performance: A new dimension for hrm," *Management decision*, vol. 46, no. 2, pp. 284–309, 2008.
- [2] M. Gjoreski, H. Gjoreski, M. Luštrek, and M. Gams, "Continuous stress detection using a wrist device: In laboratory and real life," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, ser. UbiComp '16. New York, NY, USA: ACM, 2016, pp. 1185–1193. [Online]. Available: <http://doi.acm.org/10.1145/2968219.2968306>
- [3] P. Schmidt, A. Reiss, R. Duerichen, and K. Van Laerhoven, "Wearable affect and stress recognition: A review," *arXiv preprint arXiv:1811.08854*, 2018.
- [4] E. Vildjiounaite, J. Kallio, J. Mäntyjärvi, V. Kyllönen, M. Lindholm, and G. Gimel'farb, "Unsupervised stress detection algorithm and experiments with real life data," in *Progress in Artificial Intelligence*, E. Oliveira, J. Gama, Z. Vale, and H. Lopes Cardoso, Eds. Cham: Springer International Publishing, 2017, pp. 95–107.
- [5] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. Van Laerhoven, "Introducing wesad, a multimodal dataset for wearable stress and affect detection," in *Proceedings of the 2018 on International Conference on Multimodal Interaction*. ACM, 2018, pp. 400–408.
- [6] Empatica E4, <https://www.empatica.com/e4-wristband>, 2019, accessed: 2019-05-27.
- [7] O. Incel, M. Kose, and C. Ersoy, "A review and taxonomy of activity recognition on mobile phones," *BioNanoScience*, vol. 3, no. 2, pp. 145–171, 2013.
- [8] P. Siirtola, H. Koskimäki, and J. Rönning, "Personalizing human activity recognition models using incremental learning," in *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, April 2018, pp. 627–632.