PAPER • OPEN ACCESS

Stress Detection from Multimodal Wearable Sensor Data

To cite this article: Fitri Indra Indikawati and Sri Winiarti 2020 *IOP Conf. Ser.: Mater. Sci. Eng.* **771** 012028

View the article online for updates and enhancements.

Stress Detection from Multimodal Wearable Sensor Data

Fitri Indra Indikawati[1,*], Sri Winiarti[1]

[1]Informatics Engineering, Universitas Ahmad Dahlan, Ringroad Selatan, Tamanan, Bantul, DI Yogyakarta, Indonesia.

[*]fitri.indikawati@tif.uad.ac.id, sri.winiarti@tif.uad.ac.id

Abstract. Stress can be recognised by observing changes in physiological responses on the human body. Wearable sensors for stress detection are becoming more prominent in recent years due to their functionality and non-intrusive nature. By utilising data from wearable sensors, we have developed a personalized stress detection system. Our system performs classification on stress level using multimodal data from wrist-worn device Empatica E4 wearable sensor. We implemented three different classification algorithms: Logistic Regression, Decision Tree, and Random Forest and used four-class classification conditions: baseline, stress, amusement, and meditation. By evaluating the performance of the system, we demonstrate that our system can perform the best and consistent personalized stress detection using Random Forest classifier with the accuracy of 88%-99% on 15 subjects.

1. Introduction

Stress is an interesting affective state to explore because of its effects on the human body, both physically and mentally. Long-term stress can cause another health problem to appear, ranging from mild symptoms like headaches, weight gain or weight loss, sleeping trouble, until severe one such as cardiovascular disease [1].

In the earlier days, textual information from questionnaires and interview are employed to assess people's affective state. However, this method is intrusive because it interrupts the task that is being carried or there can be bias from the interviewer. Several kinds of research have been done in affective state and stress detection, i.e., using audio-visual data [2][3], writing [4], or by observing physiological responses [5]–[11].

Stress can be detected by analysing physiological changes in the human body. When people get stressed, the sympathetic nervous system (SNS) triggers physiological responses by releasing cortisol and adrenaline. The release of these hormones can cause increased in heart rate, breathing, sweat, or muscle tension. These physiological changes can then be used to detect stress in a person.

Most stress detection system uses generalized model [6], [8], [10]. However, physiological changes for each person is different from one another. Shi et al. [9] and Smets, et al. [12] create a general and personalized model but only classifies stress into two classes.

In this study, we implemented a personalized classification by using only personal physiological data and classify affective states into four-class classification conditions: baseline, stress, amusement, and meditation. We implemented three classical machine learning algorithm for classification task: Logistic Regression, Decision Tree, and Random Forest.

The remainder of this paper is as follows: Section 2 describes the dataset, exploratory analysis of the data from Empatica E4, and classification task. Then, we discuss the experimental result of our classification system in Section 3. Finally, we conclude our study in Section 4.

2. Research Method

In this section, we describe the data preprocessing and classification process in our stress detection system.

2.1. Dataset and Exploratory Data Analysis

WESAD dataset [6] is collected from 15 subjects, consists of 13 male and 2 female participants. Each subject participates in two hours of data collection protocol with four types of conditions: baseline, amusement, stress, and meditation condition. We used these four conditions data as the classes for the classification task.

WESAD provides data from two kinds of sensors: chest-worn (RespiBAN) and wrist-worn device (Empatica E4). In this study, we use only wrist-worn device data because it is not as intrusive as a chest-worn device, and they state that wrist-worn device data is quite promising to use for classification process [6].

Empatica E4 records six raw physiological and motion data: 3-axis accelerometer (ACC), blood volume pulse (BVP), electrodermal activity (EDA), temperature (TEMP), heart rate (HR), inter-beat interval (IBI). HR and IBI is a derivative data from BVP data, and we can ignore it in the dataset. We exclude ACC data because ACC data leads to low classification score compared to results using physiological data (BVP, EDA, TEMP) [6]. We used BVP, EDA, and TEMP data as features in the classification process. We can see the details for physiological data from Empatica E4 in Table 1.

Table 1 Empatica E4 data modalities

Data	Sampling	Description
ACC	32Hz	Accelerometer data consists of 3 axis channels
BVP	64Hz	Data from a photoplethysmograph (PPG).
EDA	4Hz	Consists of a tonic (referred to as skin conductance level (SCL)) and a phasic (skin conductance response (SCR)) component.
TEMP	4Hz	Temperature in ˚C

From Table 1, we can see that each data have different sampling rate. BVP data has a higher sampling rate than EDA and TEMP data. Therefore, the number of BVP data is significantly higher than the other data. We applied downsampling to BVP data into the window size of 0.25 seconds, similar to the EDA and TEMP data.

Figure 1 shows the histogram sample for BVP (a), EDA (b), and TEMP(c) data from one of the test subject. We can see the distribution for each raw data, and there is no outlier presents in the sample data.

(a)   (b)
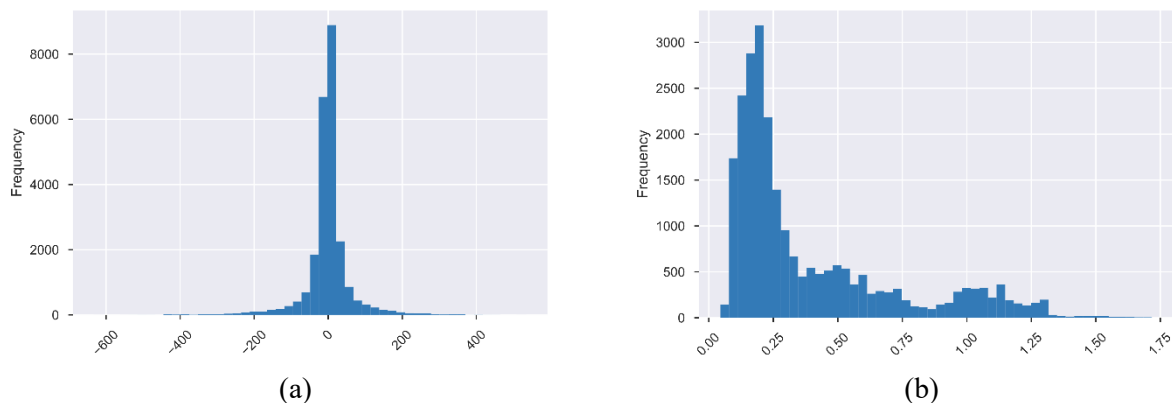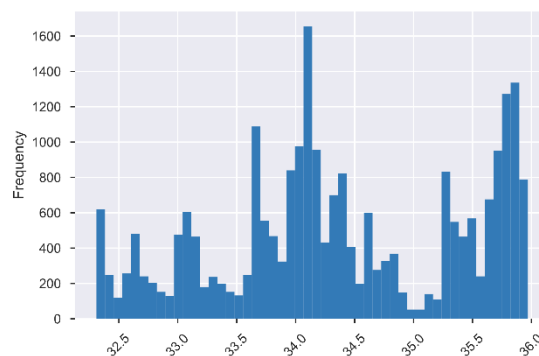
(c)

Figure 1 Histogram sample for Empatica data: (a) BVP (b) EDA (c) TEMP

2.2. Classification

The data considered in this study belong to four conditions (baseline, stress, amusement, and meditation) amount to approximately 180,000 records from a total of approximately 250,000 records or approximately 12,000 records per subject. From these records, ±40% belongs to the baseline condition, ±22% belongs to the stress condition, ±12% belongs to the amusement condition, and the remaining ±26% belongs to the meditation condition.

The experiments in WESAD [6] use binary class (stress, non-stress) and three- class (baseline, stress, and amusement) with accuracy result of 92% and 78%, respectively. Because the meditation condition records are significant, we included meditation condition as one of our four-class classification tasks. Markovic et al. performed an experiment using the dataset from the chest-based device but only use some part of the data in random [10].

Both experiments [6] [10] uses a generic model by combining all subjects data for constructing a training model. However, each person has a different degree of physiological changes in their body when faced with stress and amusement conditions. Because of this reason, we train the model for each subject separately in order to improve the classification accuracy.

We implement classic classification algorithms: Logistic Regression[13], Decision Tree[14], and Random Forest[15] using PySpark on top of Apache Spark. In our experiments, we divide data into 60% of training data and 40% of test data. We performed the experiments 10 times for each classification algorithm and each subject and randomised training and test data in each iteration.

3. Result and Discussion

In this section, we report the experimental evaluation of our classification system on a real dataset from a wearable sensor.

3.1. Experimental Setup

3.1.1. Environments. We implemented our classification system on Apache Spark [16] version 2.4.4 by utilizing PySpark (Python Spark Machine Learning Library). All experiments were conducted on a commodity machine with Intel Core i7-4500U 1.80 GHz CPU and 8 GB memory running a Windows 7 64-bit operating system. To ensure the reliability of the experiment results, we performed each classification algorithm for each subject ten times and averaged all the reported results.

3.1.2. Wearable Sensor Dataset. For evaluating our classification system, we used a real dataset named WESAD (Wearable Stress and Affect Detection) dataset [6]. The dataset is collected from two kinds of wearable sensor: RespiBAN and Empatica E4, a chest-worn and wrist-worn device, respectively. In this experiment, we only use Empatica E4 data because the authors concluded that data from the wrist-worn device is quite promising to use and because of the non-intrusive nature of the wrist-worn device.

In total, data were analysed and collected from 17 subjects. However, only data from 15 subjects are available because two subjects data has to be discarded due to sensor malfunction. WESAD dataset consists of millions of rows of raw sensor data which is collected from approximately 2 hours of the observation period. The total size of the synchronised data from the two sensors is more than 12 gigabytes.

3.2. Experiment Result
To show the performance of our classification system, we conducted experiments based on WESAD dataset using three multiclass classification algorithms: Logistic Regression, Decision Tree, and Random Forest. Table 2 shows the accuracy of the three classification algorithms on the four-class classification task: baseline, stress, amusement, and meditation.

Table 2 Accuracies result per subject on the four-class classification task

Subject	Logistic Regression (%)	Decision Tree (%)	Random Forest (%)
S2	75.89	86.69	93.08
S3	60.77	84.82	98.35
S4	73.85	79.40	91.60
S5	85.00	78.69	96.48
S6	89.81	80.41	99.57
S7	77.22	76.68	94.04
S8	62.21	94.44	99.21
S9	94.41	85.36	98.30
S10	85.41	92.65	97.02
S11	74.66	86.10	97.01
S13	66.36	74.13	88.88
S14	71.97	82.57	99.07
S15	87.73	85.05	99.76
S16	87.77	86.39	98.25
S17	88.71	87.75	99.61

We can see from Table 2 that Random Forest classifier produce the best and consistent classification accuracy compared for all of the subjects ranging from 88% to 99% classification accuracy. Compared to previous experiments which use generic model [6] [10], we can improve accuracy for our classification system by using a four-class classification problem and more personalized model by training the data per subject. Because we build our model by training the data for each subject, the accuracy result can be significantly different for a different subject. The result for S13 generally has lower accuracy for all three classification algorithms compared to the other subjects. Because each person has a different degree of physiological changes in their body therefore we cannot generalized the model for all subjects.

Figure 2 shows the confusion matrix for best prediction result (a) and worst prediction result (b) using Random Forest classifiers on the test data. In Section 2.2, we explained that we divided the dataset into train and test data, 60% and 40% respectively. Therefore we have approximately 4,000 records for test data per subject. In the best prediction scenario, the ratio for the true positives is ±40% for baseline condition, ±22% for stress condition, ±12% for amusement condition, and 26% for meditation condition. This ratio is consistent with the original dataset ratio explained in Section 2.2. False prediction happened when it tried to distinguish between baseline and meditation condition. It classifies meditation as a baseline condition. The main reason for this is because physiological data for baseline and meditation is stable compared to high physiological changes in high stimuli conditions (stress and amusement). In the worst-case scenario, it fails to classify the meditation condition. Also, it misclassified amusement condition as a baseline condition.
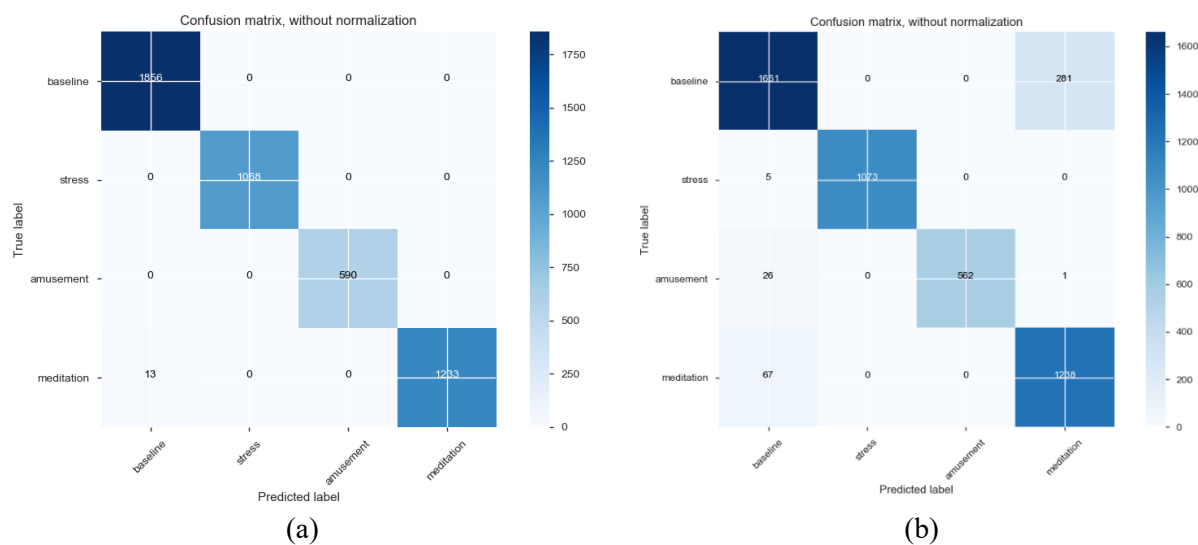
(a)   (b)

Figure 2 confusion matrix for best prediction (a) and worst prediction (b) scenario

4. Conclusion

In this paper, we have implemented a personalized stress detection system from real wearable sensor dataset. We implemented three classification algorithms to the dataset and used four-class classification: baseline, stress, amusement, and meditation. Instead of using the generalized model, we train our data for each subject because physiological changes for each person is different and we need to make a personalized model for the classification. Our classification system can make a reasonable prediction by utilising physiological data from a wrist-worn device. The best accuracy is shown when we apply Random Forest classifier into the dataset with an 88-99% accuracy. Further work is required to create a more personalized model by utilising questionnaires available in the dataset. Moreover, we can also classify different stress level (low stress, moderate stress, and high stress) and amusement level (happy, sad, angry).

5. Acknowledgement

References
[1] B. S. McEwen and E. Stellar, "Stress and the Individual: Mechanisms Leading to Disease," *Archives of Internal Medicine*, vol. 153, no. 18, pp. 2093–2101, 1993.
[2] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2017, pp. 2227–2231.
[3] P. Tzirakis, J. Zhang, and B. W. Schuller, "End-to-end speech emotion recognition using deep neural networks," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2018-April, no. 8, pp. 5089–5093, 2018.
[4] A. Gangemi, V. Presutti, and D. Reforgiato Recupero, "Frame-based detection of opinion holders and topics: A model and a tool," *IEEE Computational Intelligence Magazine*, vol. 9, no. 1, pp. 20–30, 2014.
[5] Q. Xu, T. L. Nwe, and C. Guan, "Cluster-based analysis for personalized stress evaluation using physiological signals," *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 1, pp. 275–281, 2015.
[6] P. Schmidt, A. Reiss, R. Duerichen, and K. Van Laerhoven, "Introducing WeSAD, a multimodal dataset for wearable stress and affect detection," in *ICMI 2018 - Proceedings of the 2018 International Conference on Multimodal Interaction*, 2018, pp. 400–408.

[7] N. Keshan, P. V. Parimi, and I. Bichindaritz, "Machine learning for stress detection from ECG signals in automobile drivers," in *Proceedings - 2015 IEEE International Conference on Big Data, IEEE Big Data 2015*, 2015, pp. 2661–2669.

[8] M. Gjoreski, M. Luštrek, M. Gams, and H. Gjoreski, "Monitoring stress with a wrist device using context," *Journal of Biomedical Informatics*, vol. 73, pp. 159–170, 2017.

[9] Y. Shi *et al.*, "Personalized Stress Detection from Physiological Measurements," in *Second International Symposium on Quality of Life Technology*, 2010, pp. 28–29.

[10] D. Marković, D. Vujičić, D. Stojić, Ž. Jovanović, U. Pesović, and S. Randić, "Monitoring System Based on IoT Sensor Data with Complex Event Processing and Artificial Neural Networks for Patients Stress Detection," in *2019 18th International Symposium INFOTEH-JAHORINA, INFOTEH 2019 - Proceedings*, 2019, pp. 1–6.

[11] B. Kikhia *et al.*, "Utilizing a wristband sensor to measure the stress level for people with dementia," *Sensors (Switzerland)*, vol. 16, no. 12, p. 1989, 2016.

[12] E. Smets *et al.*, "Comparison of machine learning techniques for psychophysiological stress detection," in *Communications in Computer and Information Science*, 2016, vol. 604, pp. 13–22.

[13] R. E. Wright, "Logistic regression.," 1995.

[14] J. R. Quinlan, "Induction of decision trees," *Machine learning*, vol. 1, no. 1, pp. 81–106, 1986.

[15] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[16] Apache Software Foundation, "Apache Spark." [Online]. Available: https://spark.apache.org/docs/latest/api/python/pyspark.ml.html.