April 2021

Analysing the Performance of Stress Detection Models on Consumer-Grade Wearable Devices

Van-Tu NINH [a], Sinéad SMYTH [b], Minh-Triet TRAN [c], and Cathal GURRIN [a]
[a] *School of Computing, Dublin City University*
[b] *School of Psychology, Dublin City University*
[c] *VNU-HCM, University of Science*

Abstract. Identifying stress levels can provide valuable data for mental health analytics as well as labels for annotation systems. Although much research has been conducted into stress detection models using heart rate variability at a higher cost of data collection, there is a lack of research on the potential of using low-resolution Electrodermal Activity (EDA) signals from consumer-grade wearable devices to identify stress patterns. In this paper, we concentrate on performing statistical analyses on the stress detection capability of two popular approaches of training stress detection models with stress-related biometric signals: user-dependent and user-independent models. Our research manages to show that user-dependent models are statistically more accurate for stress detection. In terms of effectiveness assessment, the balanced accuracy (BA) metric is employed to evaluate the capability of distinguishing stress and non-stress conditions of the models trained on either low-resolution or high-resolution Electrodermal Activity (EDA) signals. The results from the experiment show that training the model with (comparatively low-cost) low-resolution EDA signal does not affect the stress detection accuracy of the model significantly compared to using a high-resolution EDA signal. Our research results demonstrate the potential of attaching the user-dependent stress detection model trained on personal low-resolution EDA signal recorded to collect data in daily life to provide users with personal stress level insight and analysis.

Keywords. Stress detection using Electrodermal Activity signal, Model Selection, Statistical Analysis, Hypothesis Testing.

1. Introduction

Recently, there has been a significant volume of research into personal sensing, with various applications in self-quantification, lifelogging, and healthcare. Specifically, as wearable devices for health tracking via physiological data are becoming more popular, researchers are beginning to exploit such kinds of personal data and analyze them to provide not only insights into an individual's daily life activities and movement [6], but also meaningful insights into both physical and mental health status. For example, accelerometer and gyroscope data generated from wearable devices and smartphones can be used to recognize human activities [12, 2]. In addition, heart rate (HR) and electrodermal activity (EDA)/skin response (SR) are discriminative signals that can provide clues to per-

sonal affective states and stress levels [3]. Many approaches to build an automatic emotion state/stress level discriminator using biometric data have been proposed, but most of them investigate the performance of the detection model using professional grade, high-resolution devices in controlled laboratory settings [25, 18, 15]. Moreover, although the performance of stress detection models using Heart Rate (HR) and Heart Rate Variability (HRV) from wearable devices as well as data validation of HR signal and HRV are approved by many works [13, 19, 24, 25, 14, 9], there is limited use of low-resolution (consumer-grade) EDA signals recorded from wearable devices and little is known on the resultant effect on the performance of stress detection [23, 15, 18, 21, 8, 26]. Therefore, in this paper, we concentrate on studying and comparing different approaches of constructing stress detection models using low-resolution EDA signals.

Stress is defined as a physical, mental, or emotional response of the body to anything that requires attention or action [18] and can be categorized into three main types depending on the level of impact it can have: acute stress, episodic acute stress, and chronic stress. As chronic stress and anxiety are risk factors for dementia and cardiovascular diseases during aging [5], there is a need for some form of acute stress detection system that can monitor the stress level of an individual in daily life, either to generate a source of metadata for personal health data, or as a trigger for therapies to alleviate its harmful effects. Therefore, more work on daily-life stress detection using ubiquitous low-cost devices needs to be carried out to explore promising techniques and challenges in this area.

There are two common challenges when building a stress detection model. One of the challenges is that the stress monitoring system needs to adapt to the physiological data of every individual since people have different physiological reactions to stress according to Schmidt et al. [17], therefore, it is likely that stress detection models operate more accurately when it is customised for the individual, rather than for a population. However, no work has been done to provide a conclusion of which of these two stress detection models is an optimal in terms of proving that the conclusion does not happens by random chance. Especially, no comparison and conclusion has been made to determine whether a user-dependent or user-independent approach is better when training stress detection models on low-resolution signals. Accordingly, there is a need for a comparison between the effect of different signal-resolution on the model's stress level discrimination capability.

In summary, we present three main contributions in this paper:

- From a sample of participants' data, we show that using low-resolution EDA signal as input for the stress detection model maintains the same performance of the models as high-resolution EDA signals in the same condition of the study protocol, regardless of the choice of learning model.
- With supportive evidence, we show that stress detection models with low-resolution EDA signal input from both wrist-worn device and finger-mounted sensor manage to achieve high balanced accuracy (BA) scores ranging from 66.10% to 100% with mean and median BA of 90.53% and 93.00% respectively using Support Vector Machine models. This also shows the potential of using EDA data recorded from sensors of wearable devices to keep track of stress level.
- Additionally we show that user-dependent stress detection models are statistically more accurate than user-independent models, regardless of the choice of learning model.

2. Related Work

Heart rate (HR) and Electrodermal Activity (EDA) were found to be discriminative signals for stress level measurement [3]. Although many data validation works were performed and approved the efficiency of using HR and HRV of wearable devices in stress detection problems, there was not much research on the efficiency of EDA signal of wearable devices to construct stress detection models. Specifically, Menghini et al. showed that there is neither correlation nor visual resemblance between wrist-EDA and finger-EDA measurement and suggested more studies should be performed to assess the responsivity of wrist-EDA signal to emotional and cognitive stress [13]. Milstein and Gordon concluded that the low-resolution EDA data recorded from Empatica E4 wrist-band, which is a popular wrist-worn device for real-time physiological data streaming, is not reliable [14]. These are the motivations for us to conduct more experiments to validate the effect of low-resolution EDA signal recorded by wearable devices on the stress detection accuracy of the models.

In 2018, Schmidt et al. introduced a new publicly available dataset named WESAD for wearable stress detection and provided preliminary work on their new dataset [18]. According to their work, the Linear Discriminant Analysis (LDA) model achieved the highest accuracy score of 93.12% for binary (stress/non-stress) classification using multimodal physiological signals recorded by a chest-worn device. However, their evaluation metric using accuracy was not appropriate due to the unequal number of stress and non-stress labels in WESAD dataset which induced bias when evaluating the results. Siirtola continued this work by analysing the performance of user-independent stress detection models and the effect of window-size on recognition rates using low-resolution signals recorded by Empatica E4 wrist-band only [20]. Their best balanced accuracy score of 87.4% was achieved by training LDA model with the combination of Skin Temperature (ST), Blood Volume Pules (BVP), and HR. However, the best balanced accuracy score of the Random Forest model trained on EDA signal was only 78.3%.

For the comparison of user-dependent and user-independent models, Nkurikiyeyezu et al. conducted experiments on the performance of these two models using chest-worn HRV and EDA signals from the WESAD [18] and SWELL datasets [11] trained with a well-defined-hyperparameter Random Forest model [15]. To avoid the pitfall when choosing accuracy as an evaluation metric, they downsampled the dataset by randomly discarding samples of majority classes to balance the number of categories in the dataset. Through their experiment, the user-dependent models achieved higher accuracy than user-independent one in both datasets. They also proposed a hybrid calibrated model to improve the performance of the user-independent model from $42.5\% \pm 19.9\%$ to $95.2\% \pm 0.5\%$ by including a small number of samples of the unseen subject ($n = 100$) [15]. Additionally, Siirtola and Röning also made a comparison of user-independent and user-dependent stress detection models using the AffectiveROAD [7] dataset [21]. They suggested applying a subject-wise feature selection to improve user-independent model instead of purely building personal models using personal training data. The average balanced accuracy scores of both their user-independent and personal stress detection models were not encouraging, especially when using low-resolution EDA data only. Therefore, in our work, we propose an optimal approach of constructing high-accuracy stress detection models including feature extraction process with detailed description of the model training methods by proving the two hypotheses in section 3.1.

3. Method

3.1. Initial Hypotheses

Our experiments are conducted to validate the impact of EDA signal's quality recorded from both wrist-worn device and finger-mounted sensors on the prediction capability of stress detection models to prove that the proposed method for the stress detection model construction is trustworthy. Moreover, we evaluate the stress detection accuracy of user-independent and user-dependent models statistically using balanced accuracy metric to propose the optimal approach of training stress detection models with low-resolution EDA signal. These two tasks are completed by either proving or disproving the two following hypotheses:

- *Hypothesis 1*: Using low-resolution EDA signal, the user-dependent stress detection model is statistically more accurate than the user-independent one in terms of discriminating stress an non-stress patterns.
- *Hypothesis 2*: The stress detection accuracy of the user-dependent model trained on low-resolution EDA signal is not statistically different from the one trained on high-resolution EDA signal, which implies that the EDA signal from consumer-grade wearable devices does not deteriorate stress detection accuracy of the model.

3.2. Datasets

In our experiments, we use the WESAD [18], AffectiveROAD [7], and DCU-NVT-EXP1 datasets to prove hypothesis 1 and 2. All three datasets are used to compare the stress detection accuracy of user-independent and user-dependent models statistically, which addresses the first hypothesis. The second hypothesis is addressed using WESAD with supportive results from AffectiveROAD and DCU-NVT-EXP1 datasets as only WESAD dataset contains the records of synchronized high-sampling rate (700 Hz) and low-sampling rate data (4 Hz).

WESAD dataset: The released dataset consisted of physiological data collected from 15 participants under two different study protocols including a combination of amusement/stress/relaxation conditions. Schmidt et al. recorded high-resolution physiological Blood Volume Pulse (BVP), Electrocardiogram (ECG), Electrodermal activity (EDA), Electromyogram (EMG), Respiration (RESP), Skin temperature (TEMP), and motion (ACC) modalities from a chest-worn device named RespiBAN. Concurrently, participants were also required to wear the Empatica E4 wrist-band to capture low sampling-rate heart rate and EDA data simultaneously. The authors used study protocol as ground-truth. In our experiment, we consider the chest-worn data as WESAD-Chest and the wrist-worn data as WESAD-Wrist.

AffectiveROAD dataset: The dataset was collected to identify drivers' state indicators such as stress and arousal. Haouij et al. gathered low-resolution BVP and EDA data from an Empatica E4 wrist-band as well as high-resolution ECG and RESP data using a chest-worn device called Zephyr BioHarness 3.0. 10 participants were invited to join in 14 different driving tasks. The stress level of the participant was rated on a continuous "stress" metric ranging from 0 (no stressful) to 1 (extremely stressful) by the experimenter using a slider sitting in the rear of the car during each driving task. The subjective stress level was then validated again by the participants. In our work, we choose the

optimal threshold of 0.4 to divide the continuous "stress" metric into stress/non-stress labels as Siirtola and Röning suggested in their work. This is used as ground-truth of AffectiveROAD dataset in our experiment.

DCU-NVT-EXP1 dataset: To support our second hypotheses in 3.1 to show that the null hypothesis cannot be disproved, we created a dataset named DCU-NVT-EXP1 using finger-mounted sensor MINDFIELD eSense Skin Response (5 Hz) to record low-resolution EDA signal of participants during a study protocol containing both daily-life tasks and Virtual Reality tasks. The first stress test was the Trier Social Stress Test (TSST) [10]. The second stress test was the Sing-a-Song Stress Test (SSST) [1] and the final one was the Virtual Reality (VR) Stress Test including five stress-stimulated VR games. In total, 7 participants joined the study protocol and conducted self-evaluation forms including Subjective Units of Distress Scale (SUDS) for subjective stress evaluation and NASA Task Load Index (NASA-TLX) for an overall workload score. The subjective stress self-evaluation was then divided into four degrees of stress which were relaxed (0-20), mildly-stressed (30-50), stressed (60-80), and extremely stressed (90-100). From this division, we continued to categorize them into binary categories including non-stress (relaxed & mildly-stressed) and stress (stress & extreme stress) with the assumption that the target-user should be warned about his/her stress status when he/she experiences high degree of anxiety or distress. The binary categories of stress/non-stress status were used as the ground-truth in this dataset.

3.3. Statistical EDA Feature Extraction

The EDA signal was filtered through a 5 Hz fourth-order Butterworth low-pass filter if its Nyquist frequency is in range of 0 and 1. In the context of our experiment, this means that only the high-resolution EDA signal from WESAD-Chest data is pre-processed through a low-pass filter before extracting the feature. Then, EDA features which comprise of Skin Conduction Response (SCR), Skin Conduction Level (SCL), SCR Peaks, SCR Onsets, and SCR Amplitude were extracted. Finally, the statistical features from four works in the same field [4, 15, 18, 8] are computed except for the slope of EDA signal along the time-axis. The window-size and window-shift used in this feature extraction process are 60 seconds and 30 seconds respectively, which means that the model searches for stress pattern after each 30 seconds by observing the statistical features extracted from signals in the 60-second interval from the current signal point.

3.4. Stress Detection Model Training

Using the aforementioned statistical features extracted in section 3.3, we train stress detection models using Grid Search on five regularly used Machine Learning models which are Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), Multi-layer Perceptron (MLP), and K-Nearest Neighbors (KNN). Standardization of the feature along its axis is applied for the input of SVM, MLP, and KNN as these Machine Learning models work effectively on standardized data. We perform Leave-One-Group-Out (LOGO) strategy for the user-independent stress detection model. The training and test data for user-dependent training is divided using a stratified train-test split strategy whose test size equals to 28.6%. We are also aware of the imbalanced characteristic of the datasets when training the models by setting *class_weight* parameters in Grid Search. The Grid Search parameters for each Machine Learning model are shown in Table 1.

Table 1. Grid Search configurations of five Machine Learning models used when training stress detection models.

Model	GridCV Parameters	Values
LR	C (Regularization)	0.001, 0.01, 0.1, 0.25, 0.5, 0.75, 1, 10
	class_weight	None, balance
RF	n_estimators	500, 1000
	min_samples_split	2, 4
	min_samples_leaf	1, 4
	class_weight	None, balance
SVM	C (Regularization)	0.001, 0.01, 0.1, 0.25, 0.5, 0.75, 1, 10
	class_weight	None, balance
MLP	hidden_layer_sizes	64, 128, 256, 512
KNN	n_neighbors	3, 5, 7
	weights	uniform, distance

3.5. Statistical Analysis

Although many works show the accuracy and precision of stress detection models using low-resolution EDA signal data [15, 18, 26, 21], they only provide accuracy and precision score of detection models on specific datasets without performing statistical analyses on the results. In our experiment, we concentrate on providing conclusions based on inferential statistics, which draw conclusions about the population based on a number of representative samples. Specifically, hypothesis testing is used to estimate statistical performance of stress detection models using low-resolution EDA data recorded by available commercial wrist-worn devices or finger-mounted sensors compared to chest-worn/clinical ones based on three representative datasets as well as comparing the statistical stress detection capability between user-independent and user-dependent models.

4. Experimental Results

4.1. Evaluation Metric

As most of these datasets have unequal numbers of stress/non-stress labels, a proper evaluation metric is chosen to avoid bias assessment of experimental results. Although many pre-processing techniques can be applied to transform an imbalanced dataset into a balanced one, such as Synthetic Minority Over-sampling Technique (SMOTE) and Adaptive Synthetic Sampling Approach (ADASYNC), we do not apply these techniques to balance the dataset by removing features of dominant category since important information can be discarded through this process. Based on the analysis of Straube and Krell, balanced accuracy (BA) is an appropriate choice to evaluate results as we prefer assessing the capability of distinguishing the two categories to evaluating the precision of detecting stress patterns only [22]. Moreover, it is more convenient to use one evaluation metric in statistical analysis than using multiple evaluation scores such as accuracy, precision, recall, etc. to assess the statistical performance of a detection/prediction model with respect to one of its components or the attribute of inputs.

4.2. Statistical Analysis on the Stress Detection Capability of User-Dependent and User-Independent Models

Although the experiment made by Nkurikiyeyezu et al. concluded that a user-dependent stress detection model manages to predict more accurately than user-independent one, they only tested it on high-resolution data in the SWELL [11] and WESAD-Chest datasets [18] and trained them using a simple Random Forest model [15]. Therefore, we conduct experiments on the three datasets (WESAD-Wrist, DCU-NVT-EXP1, and AffectiveROAD) which have low-resolution EDA signal. As mentioned in section 3.1, we expect that user-dependent model detects stress patterns more accurately than user-independent one statistically, regardless of the choice of Machine Learning models.

Five user-independent and five user-dependent Machine Learning models are trained using a LOGO strategy with GridCV configurations, as described in section 3.4. Balanced accuracy scores are then computed for each individual in each dataset to create evaluation scores of both user-independent and user-dependent models respectively. Finally, hypothesis testing is conducted to either prove or disprove the initial hypothesis and provide the estimation of the difference between the population of user-dependent and user-independent Machine Learning models' performance trained on low-resolution EDA signals based on a random sample of data. The data contains $2 \times 5 \times 35 = 350$ observations in total for both outputs from user-dependent and user-independent models. Each consists of independent outputs of 5 Machine Learning models $\times 35$ participants' data in three datasets: AffectiveROAD, WESAD-Wrist, and DCU-NVT-EXP1.

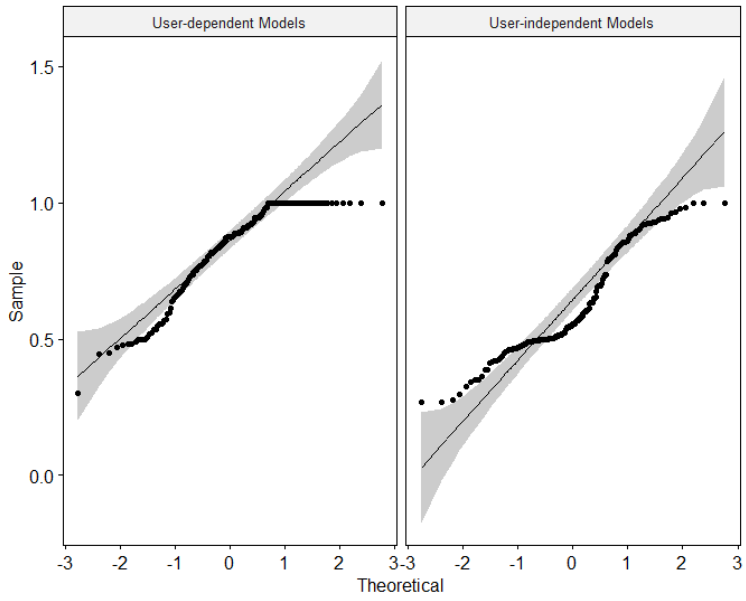

Figure 1. QQ-Plot of the performance of user-dependent and user-independent stress detection models trained on low-resolution EDA signal

Firstly, we employ Shapiro-Wilk test (SW) to test the normality of two independent data from two groups as it is proved to be the most powerful test for data normality [16].

Moreover, the Anderson-Darling test is also utilized to support the conclusion of the normality of this data. In these tests, the distribution of non-absolute difference data (BA scores of user-dependent model minus BA scores of user-independent model) is compared with the normal distribution where null hypothesis H_0 assumes that the data comes from the normal distribution. The p-values of the Shapiro-Wilk test of user-dependent and user-independent group are 2.66×10^{-8} and 1.425×10^{-15} respectively, which are both smaller than the alpha level of 0.05. In addition, the test statistics of Anderson-Darling test for both user-dependent and user-independent groups are 6.1268 and 6.3181 correspondingly, which are both larger than the critical value of 0.722 at significant level of 0.05. These values indicate that the distributions of the two sample data are not normal as the null hypothesis of both tests are rejected. This can be seen visually in Figure 1. Therefore, the Wilcoxon rank sum test is applied to prove that the user-dependent model discriminates stress and non-stress patterns statistically better than the user-independent ones. This implies that the difference of the medians between the user-dependent and user-independent models should be large and the median of the user-dependent model should be almost always in the right of the one of user-independent model. From this point, the null and alternative hypotheses of the one-sided Wilcoxon rank sum test are established as follows:

- $H_0 : M_1 = M_2$ (The user-dependent model has no improvement in distinguishing stress and non-stress patterns).
- $H_a : M_1 > M_2$ (The user-dependent model manages to distinguish stress/non-stress patterns statistically more accurately than user-independent one).

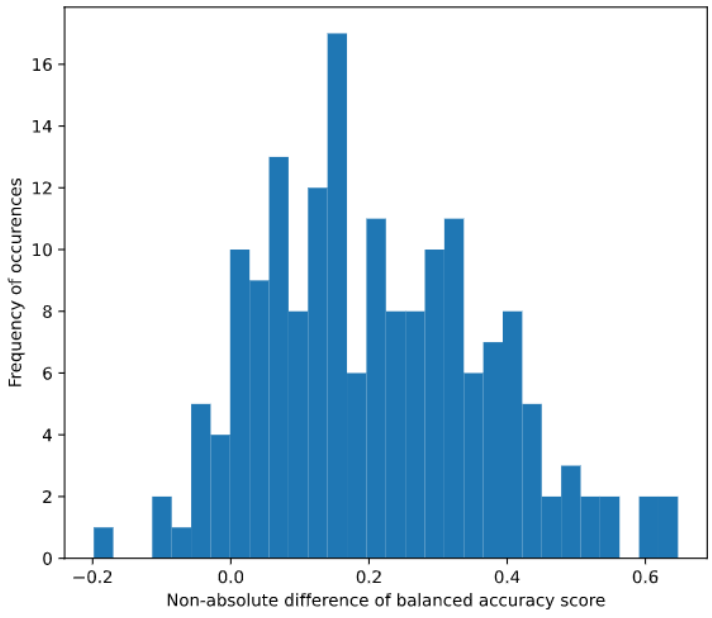

Figure 2. Distribution of non-absolute difference of balanced-accuracy scores between user-dependent and user-independent model from AffectiveROAD, WESAD-Wrist, and DCU-NVT-EXP1

In the above hypotheses, M_1 and M_2 indicate the median of stress/non-stress prediction accuracy scores of user-dependent and user-independent models respectively. The p-value of the Wilcoxon rank sum test computed from the test-statistic value of 6303 is 7.89×10^{-22}, which is significantly smaller than the pre-determined value of significance of 0.001. This implies that there is enough evidence to reject the null hypothesis with confidence level of 99.9%. This also means that statistically, the user-dependent model outperforms the user-independent one. The confidence interval ranges from $-\infty$ to -0.18. This indicates that the median of stress/non-stress prediction accuracy scores of the user-independent model is statistically less than the one of user-dependent model at most 0.18 with the confidence level of 99%. In addition, the median of the difference between two samples is estimated to be -0.2401, which is significant. From these findings, the user-dependent model can be concluded to discriminate stress and non-stress patterns statistically more accurately than the user-independent model in terms of balanced accuracy. The conclusion can be seen via Figure 2 using histogram plot.

4.3. Statistical Analysis on the Effect of Signal-resolution of EDA Signal on the Performance of Stress Detection Models

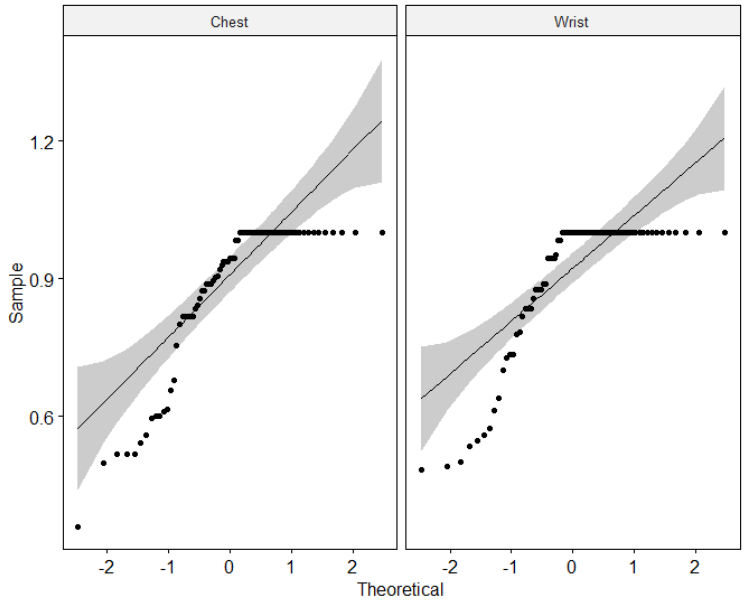

Figure 3. QQ-Plot of the performance of user-dependent stress detection models trained on WESAD-Chest and WESAD-Wrist

As the user-dependent stress detection model is proven to be more efficient than user-independent one, therefore, we only consider the evaluation scores of user-dependent model in this experiment. As WESAD dataset contains both low-resolution and high-resolution EDA signal, only this data is used in this experiment to compare the stress detection accuracy of the user-dependent model in terms of discriminating stress and non-stress patterns trained on either low-resolution EDA signal from wrist-

worn device or high-resolution one from chest-worn/clinical device regardless of different Machine Learning models. The approach of testing the hypothesis is the same as in section 4.2. Since the normality tests (WESAD-Wrist: Shapiro-Wilk's p-value of $1.09 \times 10^{-9} < 0.05$ and Anderson-Darling's test statistic of $9.985 > 0.722$, WESAD-Chest: Shapiro-Wilk's p-value of $2.04 \times 10^{-11} < 0.05$ and Anderson-Darling's test statistics of $7.003 > 0.722$) indicate that the samples from both independent groups do not follow a normal distribution with the confidence of 95%, Wilcoxon rank sum test is applied in this experiment. This can be seen visually through the QQ-plot in Figure 3. The total number of independent observations used in this research question is $2 \times 5 \times 15 = 150$, which includes 5 Machine Learning models $\times$ 15 participants $= 75$ observations. The hypotheses of the two-sided Wilcoxon rank sum test are stated as follows:

- H_0: $M_1 = M_2$ (There is no statistical difference between the stress detection accuracy of the models trained on low-resolution EDA signal and the ones trained on high-resolution EDA signal).
- H_a: $M_1 \neq M_2$ (The difference between the stress detection accuracy of these models is significant).

In the above hypotheses, M_1 and M_2 are the median of the distribution of two independent balanced accuracy scores from five machine learning models trained on WESAD-Chest and WESAD-Wrist respectively. $M_1 = M_2$ indicates that the distributions of the two data are the same, and hence have the same median. Naturally, the meaning of "the same" is estimated approximately in statistics, which infers that the difference between the two distributions via the difference of the medians is not statistically significant.

Applying the Wilcoxon rank sum test, we obtain the p-value of 0.1491, which is larger than the pre-determined value of significance of 0.05. The effect size of the test is 0.118, which is small in terms of magnitude. All of the information implies that there is not enough evidence to reject the null hypothesis with confidence level of 95%. This indicates that there is no significant statistical difference in stress detection accuracy scores when the model is trained on the features of low-resolution EDA signal. The 95% confidence interval ranges from -1.6685×10^{-2} to 6.1897×10^{-5}. This indicates that the difference between the median of the accuracy scores of the models using chest-worn EDA sample and the one using wrist-worn EDA data is in that range with the confidence level of 95%. In addition, the median of the difference between the two samples is estimated to be -2.7442×10^{-5}, which is also insignificant. Figure 4 using histogram plot also implies the conclusion visually.

4.4. Accuracy of the User-dependent Stress Detection Model trained on Low-resolution EDA Signal

Figure 5 illustrates supportive evidence for the conclusion in section 4.3. As can be seen in figure 5, the balanced accuracy scores of subject-ids in WESAD-Wrist dataset are mostly above 94%. Although two subject data trained with SVM model achieve the balanced accuracy score of 83.33% and 87.5%, these scores are also good enough for a stress detection model to be used in commercial product. On the AffectiveROAD dataset, the evaluation scores of the model range from 70.24% to 93.00% with its mean of 81.13%, which is lower than the results of WESAD-Wrist dataset although the same

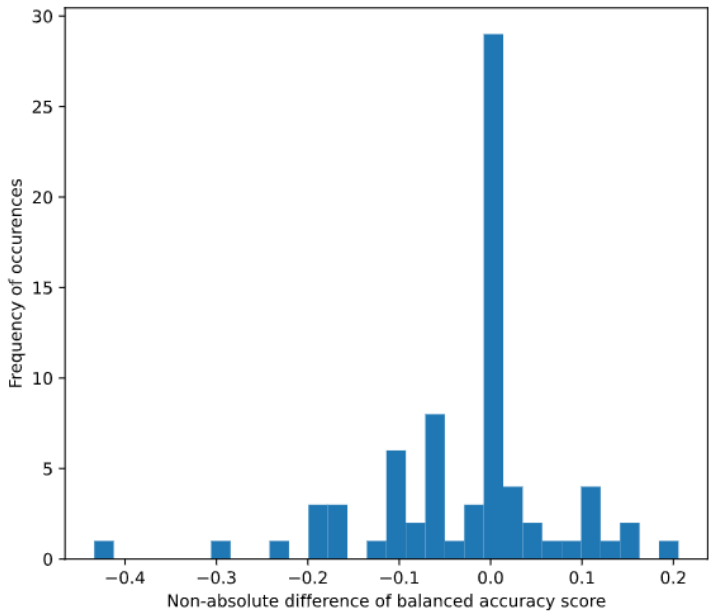

Figure 4. Distribution of non-absolute difference of balanced-accuracy scores between high-resolution and low-resolution EDA data of WESAD dataset using user-dependent stress detector.

kind of device is used to record the EDA signal. The balanced accuracy scores of subjects in DCU-NVT-EXP1 dataset are all higher than 88% with the best accuracy score that the model can achieve up to 97.73%, except for subject id 2. In summary, the mean and the median accuracy score of the user-dependent stress detection model among the three datasets using low-resolution EDA signal trained with SVM model are approximately 90.53% and 93.00% respectively, which show the potential of integrating stress detection into everyday personal data analytics so that more health insight can be obtained.

5. Conclusion

In this paper, we compare the performance of different stress detection models using EDA signal data recorded by both wrist-worn/finger-mounted devices and chest-worn device and apply inferential analysis to conclude the effect of EDA signal quality on the performance of stress detection models statistically. Through our experiment, we manage to show that with low-resolution EDA signal data, a user-dependent stress detection model is statistically provided more accurate stress pattern detection than user-independent one. In addition, we also prove that the stress detection accuracy scores of the model trained on low-resolution EDA signal has no statistically significant difference from the ones of the model trained on high-resolution one recorded by chest-worn/clinical data. The conclusion is supported by the balanced accuracy scores of stress detection SVM model trained on low-resolution EDA signal ranging from 66.10% to 100% with mean and median BA of 90.53% and 93.00% respectively. This also means

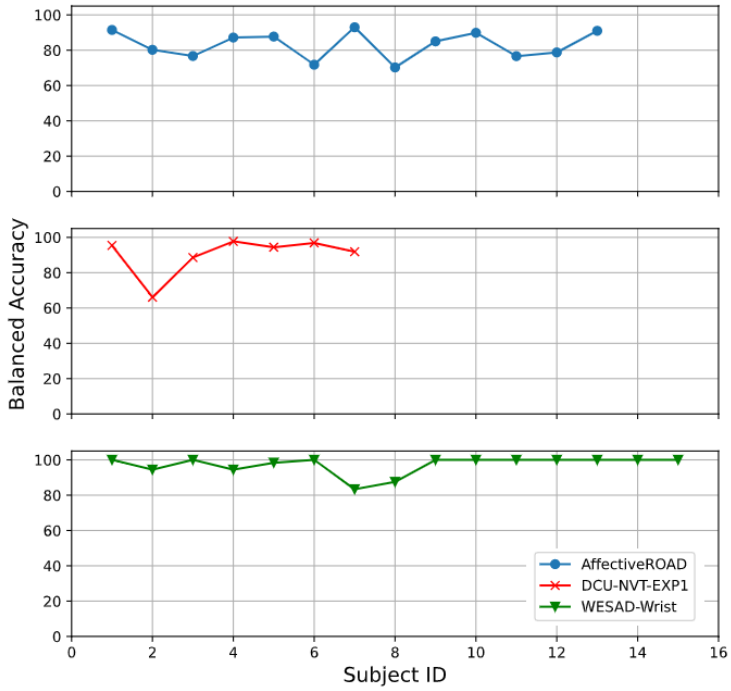

Figure 5. Balanced accuracy scores of the user-dependent stress detection model built with Support Vector Machine model for each low-resolution EDA signal dataset.

that the EDA low-resolution signal from consumer-grade devices is good enough to build user-dependent stress detection models, which makes it easier to collect data in daily life for mental health tracking and analysis.

6. Acknowledgements

This research was conducted with the financial support of ADAPT Core under Grant Agreement No. 13/RC/2106 at the ADAPT SFI Research Centre at Dublin City University. The ADAPT SFI Centre for Digital Content Technology is funded by Science Foundation Ireland through the SFI Research Centres Programme and is co-funded under the European Regional Development Fund (ERDF) through Grant Number 13/RC/2106_P2.

References

[1] The short sing-a-song stress test: A practical and valid test of autonomic responses induced by social-evaluative stress. *Autonomic Neuroscience*, 224:102612, 2020. ISSN 1566-0702. doi: https://doi.org/10.1016/j.autneu.2019.102612.

[2] Ganapati Bhat, Ranadeep Deb, Vatika Vardhan Chaurasia, H. Shill, and Ümit Y. Ogras. Online human activity recognition using low-power wearable devices. *2018*

IEEE/ACM International Conference on Computer-Aided Design (ICCAD), pages 1–8, 2018.

[3] Yekta Said Can, Niaz Chalabianloo, Deniz Ekiz, and Cem Ersoy. Continuous stress detection using wearable sensors in real life: Algorithmic programming contest case study. *Sensors (Basel, Switzerland)*, 19, 2019.

[4] Jongyoon Choi, B. Ahmed, and R. Gutierrez-Osuna. Development and evaluation of an ambulatory stress monitor based on wearable sensors. *IEEE Transactions on Information Technology in Biomedicine*, 16:279–286, 2012.

[5] Amy Gimson, Marco Schlosser, Jonathan Huntley, and Natalie L. Marchant. Support for midlife anxiety diagnosis as an independent risk factor for dementia: a systematic review. *BMJ Open*, 8, 2018.

[6] C. Gurrin, Hideo Joho, F. Hopfgartner, L. Zhou, and Rami Albatal. Overview of ntcir-12 lifelog task. In *NTCIR*, 2016.

[7] Neska El Haouij, Jean-Michel Poggi, Sylvie Sevestre-Ghalila, Raja Ghozi, and Mériem Jaïdane. Affectiveroad system and database to assess driver's attention. New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450351911.

[8] Jennifer Healey and Rosalind W. Picard. Detecting stress during real-world driving tasks using physiological sensors. *IEEE Transactions on Intelligent Transportation Systems*, 6:156–166, 2005.

[9] Hye-Geum Kim, Eun-Jin Cheon, Dai seg Bai, Y. Lee, and Bon-Hoon Koo. Stress and heart rate variability: A meta-analysis and review of the literature. *Psychiatry Investigation*, 15:235 – 245, 2018.

[10] Clemens Kirschbaum, Karl-Martin Pirke, and Dirk Hellhammer. The 'trier social stress test' – a tool for investigating psychobiological stress responses in a laboratory setting. *Neuropsychobiology*, 28:76–81, 02 1993. doi: 10.1159/000119004.

[11] Saskia Koldijk, Maya Sappelli, Suzan Verberne, Mark A. Neerincx, and Wessel Kraaij. The swell knowledge work dataset for stress and user modeling research. New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450328852.

[12] T. H. Le, Q. H. Tran, and T. L. Le. Systematic evaluation of deep learning models for human activity recognition using accelerometer. In *2020 7th NAFOSTED Conference on Information and Computer Science (NICS)*, pages 152–157, 2020. doi: 10.1109/NICS51282.2020.9335853.

[13] L. Menghini, Evelyn Gianfranchi, Nicola Cellini, E. Patron, M. Tagliabue, and M. Sarlo. Stressing the accuracy: Wrist-worn wearable sensor validation over different conditions. *Psychophysiology*, page e13441, 2019.

[14] Nir Milstein and Ilanit Gordon. Validating measures of electrodermal activity and heart rate variability derived from the empatica e4 utilized in research settings that involve interactive dyadic states. *Frontiers in Behavioral Neuroscience*, 14:148, 2020.

[15] Kizito Nkurikiyeyezu, Anna Yokokubo, and Guillaume Lopez. Effect of person-specific biometrics in improving generic stress predictive models. *Sensors and Materials*, 32:703–722, 02 2020. doi: 10.18494/SAM.2020.2650.

[16] Nornadiah Razali and Yap Bee Wah. Power comparisons of shapiro–wilk, kolmogorov–smirnov, lilliefors and anderson–darling testsg and. *Journal of Statistical Modeling and Analytics*, 2(1):21–33, 2011.

[17] Philip Schmidt, Attila Reiss, Robert Dürichen, and Kristof Van Laerhoven. Wearable affect and stress recognition: A review. *ArXiv*, abs/1811.08854, 2018.

[18] Philip Schmidt, Attila Reiss, Robert Dürichen, Claus Marberger, and Kristof Van Laerhoven. Introducing wesad, a multimodal dataset for wearable stress and affect detection. *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, 2018.

[19] Angela A. T. Schuurmans, Peter de Looff, K. S. Nijhof, Catarina Rosada, R. Scholte, A. Popma, and R. Otten. Validity of the empatica e4 wristband to measure heart rate variability (hrv) parameters: a comparison to electrocardiography (ecg). *Journal of Medical Systems*, 44, 2020.

[20] Pekka Siirtola. Continuous stress detection using the sensors of commercial smartwatch. *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*, 2019.

[21] Pekka Siirtola and Juha Röning. Comparison of regression and classification models for user-independent and personal stress detection. *Sensors (Basel, Switzerland)*, 20, 2020.

[22] Sirko Straube and Mario M. Krell. How to evaluate an agent's behavior to infrequent events?—reliable performance estimation insensitive to class distribution. *Frontiers in Computational Neuroscience*, 8:43, 2014.

[23] Roberto Sánchez-Reolid, María López, and Antonio Fernández-Caballero. Machine learning for stress detection from electrodermal activity: A scoping review, 2020. URL `https://doi.org/10.20944/preprints202011.0043.v1`.

[24] Hendrika G. van Lier, M. Pieterse, A. Garde, M. Postel, H. D. de Haan, M. Vollenbroek-Hutten, J. M. Schraagen, and M. Noordzij. A standardized validity assessment protocol for physiological signals from wearable technology: Methodological underpinnings and an application to the e4 biosensor. *Behavior Research Methods*, 52:607 – 629, 2019.

[25] Robert Wang, Gordon Blackburn, Milind Desai, Dermot Phelan, Lauren Gillinov, Penny Houghtaling, and Marc A. Gillinov. Accuracy of wrist-worn heart rate monitors. *JAMA Cardiology*, 2:104–106, 2017.

[26] Roberto Zangroniz, Arturo Martinez Rodrigo, José Manuel Pastor García, María López Bonal, and Antonio Fernández-Caballero. Electrodermal activity sensor for classification of calm/distress condition. *Sensors*, 17:2324, 10 2017. doi: 10.3390/s17102324.