Detecting Anxiety Trends Using Wearable Sensor Data in Real-World Situations

Marissa Gray[1], Shweta Majumder[1], Kate Nelson[1], and Reshma Munbodh[2,3](✉)

[1] School of Engineering, Brown University, Providence, RI, USA
[2] Department of Radiation Oncology, Alpert Medical School of Brown University, Providence, RI, USA
[3] Department of Radiation Oncology, Columbia University Irving Medical Center, New York, NY, USA
reshma.munbodh@columbia.edu

Abstract. Anxiety can manifest through a range of physiological changes. We develop methods to detect anxiety among medical residents making case presentations in a clinical setting using wearable sensors and machine learning. A number of classifiers are tested on different features extracted from real-time physiological measurements. Our results indicate that anxiety can be detected among healthy volunteers in clinical setting and serve as an introduction to future wearable sensing studies for applications in radiation oncology.

Keywords: Wearable sensors · Machine learning · Anxiety detection · Radiation oncology · Heart Rate Variability (HRV)

1 Introduction

It is well known that different mental states can manifest themselves through physiological signals. Particularly, anxiety and stress, which are the body's adaptive neural responses to threats, have been attributed to physiological changes in the body [1]. Our brain detects certain situations to be stressful including loss of control, unpredictability and level of threat to self [2]. After the brain processes a situation as "stressful," it triggers a pathway that ultimately activates the sympathetic nervous system (SNS), which leads to various physiological and behavioral changes that are detectable and distinguishable from baseline (rested) physiology [1–3]. Thus, physiological signals can be objective biomarkers to detect anxiety and stress. Heart rate (HR), heart rate variability (HRV), skin conductance or electrodermal activity (EDA)[1], and body movement or accelerometry (ACC), have been correlated successfully with stress and anxiety in a number of

[1] The variation in electrical conductance of the skin due to sweat secretion.

M. Gray and S. Majumder—These authors contributed equally to this work.

J. Bowles et al. (Eds.): DataMod 2021, LNCS 13268, pp. 107–117, 2022.
https://doi.org/10.1007/978-3-031-16011-0_8

situations including public speaking, mental health monitoring, and on-the-job anxiety [4–7].

Although the SNS is responsible for the body's physical responses to stressors, it works in tandem with the parasympathetic nervous system (PNS), which responds to rest and relaxation. These complimentary systems are highly involved in activities of the heart. Increases in sympathetic activity relate to an increased heart rate whereas heart rate decreases can be attributed to relative increases in parasympathetic activity. This dynamic balance can be altered due to different neurological or psychological activities [12]. Thus, it is not surprising that anxiety can have measurable impacts on the heart rate activity [7].

Public speaking, an essential skill for learning and career development, is also a task that contributes to social stress and anxiety among many individuals. The National Institute of Mental Health states that around 73% of the U.S. population has a fear of public speaking even though it is a common occurrence in everyday life [26]. For many, it is a regular part of their professional careers. Because there is a high prevalence of public speaking anxiety, many researchers have used wearable sensors to detect physiological features of anxiety during public speaking tasks. However, most of these experiments have been held in laboratory settings with controlled tests on subjects [9]. Although there have been studies focused on targeted groups of people in real-world settings, they have often been limited in scope and to controlled activities. Few studies have focused on real-time anxiety detection in real-world activities [4,5]. Here, we present a feasibility study to detect anxiety throughout real-world public speaking situations that are an essential job function. Specifically, we address the detection of anxiety among medical residents making case presentations in a hospital-based radiation oncology department.

2 Related Work

Previously, wearable sensors have been used to detect physiological signals of stress and anxiety which are then used to train classifier algorithms that aim to quantify stressful or anxious states [8,9]. Studies analyzing HRV, EDA [9,10], and ACC [11] have been conducted in a wide range of settings. However, most of these studies are carefully controlled with participants undergoing specific public speaking tests such as the Trier Social Stress Test [24] and the Maastricht Acute Stress Test [25]. In these studies, physiological signals are analyzed using a variety of supervised learning techniques such as support vector machines (SVM), decision trees, k-Nearest Neighbors (KNN), random forests, logistic regression, and multilayer perceptron [22,23,26]. The acquired physiological signals vary between studies. Several studies acquire multimodal sensor information such as electrocardiography (ECG), electroencephalography (EEG), photoplethymography (PPG), and electrodermal activity (EDA) [25,26], while others focus on a single physiological signal. Lee et al. conducted a study that analyzed only EDA data from student volunteers in a real-world classroom setting [4]. The EDA data was collected from students while they were delivering regular presentations in class. After processing, the EDA data was analyzed using k-means clustering to classify data into two groups based on anxiety levels. Although this study collected physiological data in a real-world setting, researchers focused on one type of signal and a single classification method.

Heart activity is measured by looking at changes in the volume of blood at a localized area of the body, such as the wrist and is commonly detected by a PPG sensor [13,14]. PPG uses light to measure variations of blood volume. Several features from time series and frequency domain data of PPG data can be extracted and derived. These features, such as HRV, have been shown to be indicators of heart activity and can be used to gain insight into the person's psychological state [1].

In this study, we use wearable sensors to collect real-time HRV data from healthy subjects who are performing an essential job function that involves public speaking to a group of supervisors. Supervised machine learning techniques are applied to the HRV data with the aim of identifying how to best classify and predict anxiety in the subjects. Here we are testing multiple classification methods of one type of physiological signal (HRV). Our overall goal is to use this knowledge to develop real-time methods of detecting and analyzing anxiety experienced in real-world situations.

3 Methods

Healthy subjects (N = 3) were recruited from a group of radiation oncology medical residents. The subjects were studied during a known anxiety-inducing activity in medical residents: presentations during chart rounds [15–17]. During these regular presentations, residents sit and review patients' records with peers and supervising physicians. To acquire data, the subjects were fitted with an Empatica E4[2] wearable sensor on the wrist. The Empatica E4 acquires blood volume pulse (BVP), EDA, accelerometry data (ACC), and body temperature (BT). These signals were recorded for approximately 20 min before, during chart rounds (up to one hour) and after chart rounds (at least 5 min). The residents were seated during the chart rounds and their presentations. They were instructed to remain as still as possible so that the physiological measures would not be affected by motion. Deidentified physiological data was recorded in real-time and transferred to the cloud-based storage system. Residents were asked to mark the start and end of chart rounds in addition to the time they presented their cases during chart rounds using the event marking capability on the Empatica E4. Thus, the 3 main sections of the data are the baseline period (starting 20 min before chart rounds), the chart round period (lasting up to 1 h), and the recovery period (at least 5 min after chart rounds end). Each subject was involved in 2–3 chart rounds sessions. Hence, we collected a total of seven sets of data during seven different chart round sessions from the three subjects.

[2] https://www.empatica.com/en-gb/research/e4/.

3.1 Extraction of HRV Features

In this study, we use time-domain and frequency-domain HRV features to detect anxiety. Extraction of the HRV features is described below.

In the Empatica E4, heart activity is measured through BVP at a sampling rate 64 Hz. The sensor has a built-in algorithm that calculates and displays a time series of inter-beat intervals (IBIs), which is the time between two consecutive heart beats, in real-time. The IBI conversion algorithm also removes spurious peaks that can occur due to rigorous movement, and thus can lead to data loss during high intensity movements [18]. From the time series of IBIs, the instantaneous heart rate (HR) in beats per minute (bpm) is given by Eq. 1.

$$HR = \frac{60}{IBI}\, bpm \tag{1}$$

In order to account for variations in inherent heart rate among study subjects, the heart rate time series for each experimental session was normalized using the mean of the baseline heart rate data from the experiment. For the normalization, the mean of the truncated baseline data is taken by excluding 90 s of data from both the start and the end of the baseline section (see the red section of Fig. 1). The first 90 s of the data was excluded because the raw BVP signal required some time to stabilize after the device is turned on due to signal noise observed within the first few minutes of data collection.

A 90-s window with 75% overlap was used to extract HRV features from the IBI, interpolated IBI (IIBI) and HR time series. The 75% overlap was chosen to have sufficient meaningful data points for HRV feature extraction even with the short time window [9]. The time-domain HRV features extracted for each window were:

- SDNN: Standard deviation of the inter-beat intervals
- nAVHR: Normalized average heart rate
- nMAXHR: Normalized maximum heart rate
- nMINHR: Normalized minimum heart rate
- nPNN90: Normalized 90-th percentile heart rate
- NN50: The number of IBIs in a time window which differ by more than 50ms from their preceding IBI
- PNN50: The proportion of the IBIs which differ by more than 50ms from their preceding IBI in the time window.

From the IIBI time-series, a Welch's periodogram is computed [2]. Welch's method for periodogram estimation [19] uses fast Fourier transforms (FFT) on overlapping segments of the time series and averages the FFT power to estimate the power spectrum for the signal. Frequency domain HRV features extracted for each time window were:

- VLF: Band-power of the very low frequency (<0.04 Hz).
- LF: Band-power of the low frequency (0.04 Hz–0.15 Hz).
- HF: Band-power of the high frequency (0.15 Hz–0.4 Hz)
- HF/LF: The ratio of high frequency to low frequency.

3.2 Data Labeling

For data labeling, features from only the truncated baseline are were labeled as '0'. From the chart rounds data, features from 90 s before the start of the presentation to 3–5 min after the start of the presentation were labeled as '1' (the time duration depended on each session). In general, the presentation sessions lasted from 3 to 15 min. An example of the segments used to label the training data is shown in Fig. 1.

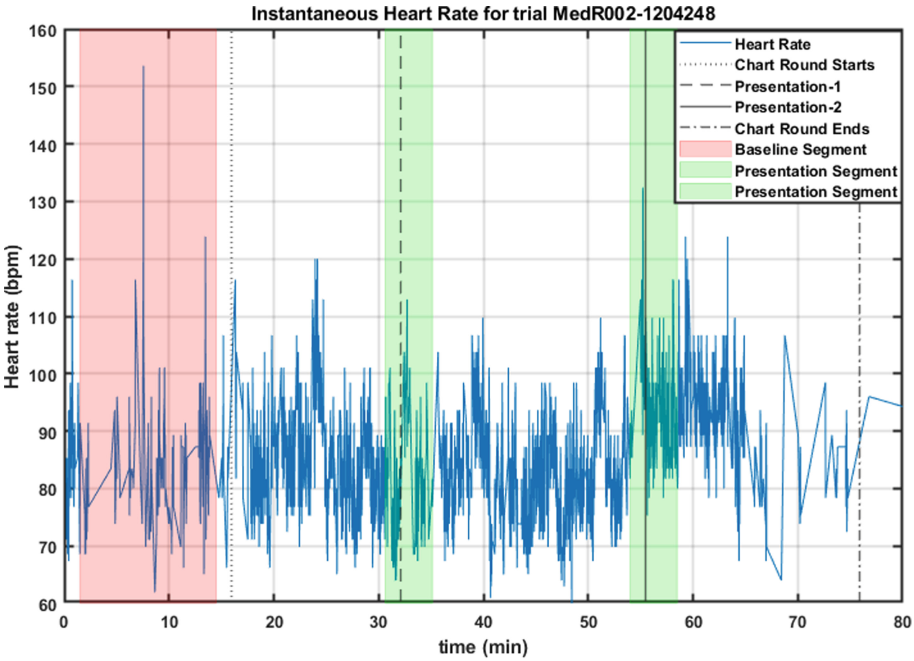

Fig. 1. How the dataset is segmented shown on the IBI time series (blue plot). Black vertical lines from left to right: i) Chart round starts, ii) Presentation 1, iii) Presentation 2, iii) Chart Round ends. Segments: i) Red: Baseline segment, ii) Green: Presentation segment (Color figure online)

3.3 Classifier Training and Validation

The labeled features (0 = not anxious and 1 = anxious) from all the first chart round sessions of each subject (N = 3) were combined into a training set. The training set was fed into the different classification algorithms with 10-fold cross validation in the MATLAB Classification Learner App and the prediction accuracy on the training sets was recorded.

Feature Selection: Three different feature selection methods were used: (1) Principal component analysis (PCA), (2) Least absolute shrinkage and selection operator (Lasso) regression and (3) the Chi-squared test.

PCA is an unsupervised dimension reduction method used to derive a low-dimensional feature set from a large number of variables [27]. The lower dimensional feature set is obtained by projecting the multivariate data points onto the first few principal components, which are selected to maximize the variance of the projected data. These projections or linear combinations of the variables constitute the predictors in the classifiers. We used PCA with 3 component retention.

Lasso regression performs both regularization and variable selection towards obtaining a classifier with improved prediction accuracy and interpretability [27]. The Lasso regularization for generalized linear models selected the features NN50, MAXHR, MINHR, VLFP, HFP.

The **Chi-squared test** is used to determine if a feature variable is independent of the response variable (i.e., the label), wherein a smaller p-value indicates that the predictor is dependent on the response variable and therefore is important. The predictor importance score calculated by the MATLAB function gives a value calculated by $-\log(p)$. Hence higher scores will indicate higher importance. Hence, the predictor importance of the 11 predictor features from IBIs was calculated using a Chi-squared test with MATLAB's built-in predictor importance function (fscchi2) [20,21] was calculated and the three features selected were AVHR, PNN90, MAXHR.

Feature Classification: The classification algorithms evaluated were: multivariate logistic regression (LR), SVM and KNN. The feature combinations/feature selection methods that were used for the classifier training were:

- FS1: PCA with 3 component retention
- FS2: Lasso GLM selected features (NN50, MAXHR, MINHR, VLFP, HFP).
- FS3: The three features with the highest predictor importance in the Chi-squred test (AVHR, PNN90, MAXHR).

3.4 Classifier Testing

The three trained classifier models (LR, SVM and KNN) were then used to predict anxiety on labeled features from the new and repeated chart round sessions for the participants (4 sessions). These test data were labeled similarly as the training data (see Sect. 3.2). Using the predicted labels from each trained classifier models on test data, classifier performances metrics such as the testing accuracy, recall and precision were calculated to compare the model performances.

4 Results

In general, presenters are aware that it is their turn to speak about 20–30 s before the start of their presentation. Verbal accounts from the residents indicated that they were aware that it was their turn to speak about 20–30 s before the start of their presentation.

As mentioned in Sect. 3.3, the labeled features from the first chart round session for each participant were used for training, and validation of different classification algorithms with the different feature selection methods. Testing was done on the remaining chart round sessions to verify testing accuracy, recall, precision and F-measure on new sets of experiments. Two-dimensional grouped scatter plots were generated of the different features extracted from IBI and HR in all the experiments. Figure 2 shows a graph of the standard deviation of the IBIs plotted against the average normalized heart rate for the baseline period and the period surrounding the presentations.

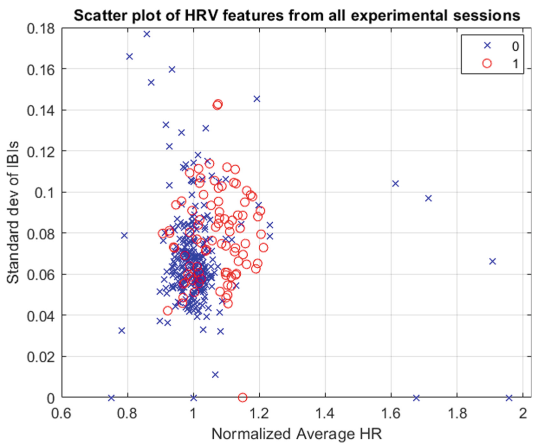

Fig. 2. Comparison of standard deviation of the IBIs with the average normalized heart rate in all 7 experimental sessions. 0 = Baseline/“Not Anxious” in Red, 1 = Presenting/“Anxious” in Blue (Color figure online)

The accuracy (on both test and training data), and precision, recall and F-measure (on test data only) of each of the classifiers with different feature combinations are shown in Table 1.

Table 1. Performance of different classifiers on different feature combinations. The accuracy (acc.), recall, precision (prec.) and F-measure (F-meas.) for SVM, KNN and LR are shown for Feature set 1 (FS1), Feature set 2 (FS2), and Feature set 3 (FS3)

Features	Classifier	Training	Testing			
		Acc.	Acc.	Recall	Prec.	F-Meas.
FS1	SVM	86.4	58.2	0.02	0.14	0.03
	KNN	89.1	71.2	0.41	0.73	0.52
	LR	86.4	62.1	0.03	0.67	0.06
FS2	SVM	90.9	61.4	0.32	0.5	0.39
	KNN	94.1	63.4	0.22	0.56	0.31
	LR	87.7	54.9	0.08	0.25	0.13
FS3	SVM	92.7	73.2	0.51	0.71	0.59
	KNN	93.6	69.9	0.44	0.67	0.53
	LR	86.8	67.3	0.15	1	0.26

5 Discussion

The main goal of this study was to identify classifiable differences in heart activity data, indicative of anxiety, in medical residents presenting to an audience. Eleven features were extracted from heart activity data. Different feature combinations were used to train supervised learning models. Preliminary analysis showed classifiable differences between features of the baseline and presentation data, with an increased normalized heart rate during the presentation period. In the absence of other known factors causing these physiological changes, the increased individualised heart rate may indicate heightened anxiety and is consistent with previous findings that have concluded that anxiety can increase heart rate from the baseline [1,4]. In this study, normalization was used to account for the inter-subject variations in heart rate. The 90^{th} percentile normalized heart rate was not used in other similar papers, but it ranked as second most important feature in our predictor importance analysis using Chi-squared test.

A specific challenge for this study is the lack of set thresholds for heart rate and other heart activity metrics that indicate anxiety. Therefore, we need to compare the heart rate data at baseline (rest) to heart rate during the anxiety-inducing situation. Modeling situational anxiety in a real-life scenario can bring about its own challenges: the lack of a controlled environment can corrupt data through movements and environmental changes. In this study, the labeling of the data was particularly challenging and it was necessary to make certain assumptions as to when the subjects were anxious.

A unique aspect of the data labeling employed in the project was to label only a fixed-time segment around the presentation as "anxious" rather than the whole presentation segment since the subjects might not stay anxious throughout chart rounds. Model accuracy on training data was highest when all HRV features were used to train an optimizable KNN. However, the high number of features in this model may indicate overfitting of the data. In this study, KNN on FS3 had the highest accuracy, recall, precision and F-score on the test data. Interestingly, SVM was the classifier of choice in similar studies in the literature. A limitation of this study is the low sample size (n = 7). To increase the number of data points from each experiment, a 75% overlap was used with the 90-s window for feature extraction, which is slightly higher than the recommended overlap.

From the initial results, we can see that there is distinguishable differences between HRV features between the baseline state and early presentation state. Also, KNN worked the best for all feature selection methods with the best testing accuracy with PCA.

This early study served as a feasibility study and an introduction to future wearable sensing studies for applications in real-world settings. In the future, as we increase the number of subjects in the study, we plan to investigate different machine learning methods to construct the classifiers including alternate schemes for splitting the data into training and test sets. We are also exploring other machine learning methods such as unsupervised learning methods which automatically group the data into distinct clusters. We plan to extend this study to include EDA and accelerometry data in the future.

Acknowledgment. We would like to thank Dr. David Wazer, Dr. Elana Nack, Dr. Imran Chowdhury and Dr. Jeff Huang for their support.

References

1. Thayer, J.F., Åhs, F., Fredrikson, M., Sollers, J.J., Wager, T.D.: A meta-analysis of heart rate variability and neuroimaging studies: implications for heart rate variability as a marker of stress and health. Neurosci. Biobehav. Rev. **36**(2), 747–756 (2012). https://doi.org/10.1016/j.neubiorev.2011.11.009
2. Ollander, S., Godin, C., Campagne, A., Charbonnier, S.: A comparison of wearable and stationary sensors for stress detection. In: 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Budapest, Hungary, pp. 004362–004366, October 2016. https://doi.org/10.1109/SMC.2016.7844917
3. Understanding the stress response - Chronic activation of this survival mechanism impairs health. Harvard Health Online, March 2011
4. Lee, H., Kleinsmith, A.: Public speaking anxiety in a real classroom: towards developing a reflection system. In: Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems, Glasgow, Scotland, UK, pp. 1–6, May 2019. https://doi.org/10.1145/3290607.3312875
5. Choi, B., Jebelli, H., Lee, S.: Feasibility analysis of electrodermal activity (EDA) acquired from wearable sensors to assess construction workers' perceived risk. Saf. Sci. **115**, 110–120 (2019). https://doi.org/10.1016/j.ssci.2019.01.022
6. Oskooei, A., Chau, S.M., Weiss, J., Sridhar, A., Martínez, M.R., Michel, B.: DeStress: deep learning for unsupervised identification of mental stress in firefighters from heart-rate variability (HRV) data. arXiv:1911.13213 [cs, eess, stat], November 2019. Accessed 10 Nov 2020
7. Fuentes-García, J.P., Clemente-Suárez, V.J., Marazuela-Martínez, M.Á., Tornero-Aguilera, J.F., Villafaina, S.: Impact of real and simulated flights on psychophysiological response of military pilots. IJERPH **18**(2), 787 (2021). https://doi.org/10.3390/ijerph18020787

8. Carreiro, S., Chintha, K.K., Shrestha, S., Chapman, B., Smelson, D., Indic, P.: Wearable sensor-based detection of stress and craving in patients during treatment for substance use disorder: a mixed methods pilot study. Drug Alcohol Depend. **209**, 107929 (2020). https://doi.org/10.1016/j.drugalcdep.2020.107929
9. Delmastro, F., Martino, F.D., Dolciotti, C.: Cognitive training and stress detection in MCI frail older people through wearable sensors and machine learning. IEEE Access **8**, 65573–65590 (2020). https://doi.org/10.1109/ACCESS.2020.2985301
10. Can, Y.S., Arnrich, B., Ersoy, C.: Stress detection in daily life scenarios using smart phones and wearable sensors: a survey. J. Biomed. Inform. **92**, 103139 (2019). https://doi.org/10.1016/j.jbi.2019.103139
11. McGinnis, R.S., et al.: Rapid anxiety and depression diagnosis in young children enabled by wearable sensors and machine learning. In: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Honolulu, HI, pp. 3983–3986, July 2018. https://doi.org/10.1109/EMBC.2018.8513327
12. Shaffer, F., Ginsberg, J.P.: An overview of heart rate variability metrics and norms. Front. Public Health **5**, 258 (2017). https://doi.org/10.3389/fpubh.2017.00258
13. "Empatica Support," E4 Wristband Data, 23 January 2020. https://support.empatica.com/hc/en-us/sections/200582445-E4-wristband-data
14. Jones, D.: The Blood Volume Pulse - Biofeedback Basics, 10 May 2018. https://www.biofeedback-tech.com/articles/2016/3/24/the-blood-volume-pulse-biofeedback-basics
15. Johnson, D.: We Are Expected To Be Public Speakers, But Never Taught, 02 April 2019. https://opmed.doximity.com/articles/we-are-expected-to-be-public-speakers-but-never-taught?csrfattempted=yes
16. Kavanagh, J.: Giving presentations without palpitations. BMJ **332**(7555), s242–s243 (2006). https://doi.org/10.1136/bmj.332.7555.s242
17. Tejwani, V., Ha, D., Isada, C.: Observations: public speaking anxiety in graduate medical education-a matter of interpersonal and communication skills? J. Grad. Med. Educ. **8**(1), 111 (2016). https://doi.org/10.4300/JGME-D-15-00500.1
18. "Empatica Support," Utilizing the PPG/BVP signal, 23 January 2020. https://support.empatica.com/hc/en-us/sections/200582445-E4-wristband-data
19. Welch, P.: The use of fast Fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. IEEE Trans. Audio Electroacoust. **15**(2), 70–73 (1967). https://doi.org/10.1109/TAU.1967.1161901
20. Univariate feature ranking for classification using chi-square tests. MATLAB (2021)
21. Gajawada, S.: "Towards Data Science," Chi-Square Test for Feature Selection in Machine learning. https://towardsdatascience.com/chi-square-test-for-feature-selection-in-machine-learning-206b1f0b8223
22. Gjoreski, M., Luštrek, M., Gams, M., Gjoreski, H.: Monitoring stress with a wrist device using context. J. Biomed. Inform. **73**, 159–170 (2017). https://doi.org/10.1016/j.jbi.2017.08.006
23. Shaukat-Jali, R., van Zalk, N., Boyle, D.E.: Detecting subclinical social anxiety using physiological data from a wrist-worn wearable: small scale feasibility study. JMIR Formative Res. **5**(10) (2021). https://doi.org/10.2196/32656
24. Mozos, O.M., Sandulescu, V., Andrews, S., et al.: Stress detection using wearable physiological and sociometric sensors. Int. J. Neural Syst. 1–17 (2005). https://core.ac.uk/reader/76958656?utmsource=linkout

25. Betti, S., Lova, R.M., Rovine, E., et al.: Evaluation of an integrated system of wearable physiological sensors for stress monitoring in working environments. IEEE Trans. Biomed. Eng. **65**(8), 1748–1758 (2018). https://doi.org/10.1109/TBME.2017.2764507
26. Arsalan, A., Majid, M.: Human stress classification during public speaking using physiological signals. Comput. Biol. Med. **133**, 10437 (2021). https://doi.org/10.1016/j.compbiomed.2021.104377
27. James, G., Witten, D., Hastie, T., Tibshirani, R.: An Introduction to Statistical Learning: with Applications in R, 2nd edn., pp. 227–250. Springer, New York (2022). https://doi.org/10.1007/978-1-4614-7138-7