

EEG-based multi-level stress classification with and without smoothing filter

Eduardo Perez-Valero^{a,b}, Miguel A. Lopez-Gordo^{b,c,d,*}, Miguel A. Vaquero-Blasco^{b,c}

^a Department of Computer Architecture and Technology, University of Granada, Spain

^b Research Centre for Information and Communications Technologies, University of Granada, Spain

^c Department of Signal Theory, Telematics and Communications, University of Granada, Spain

^d Nicolo Association, Churriana de la Vega, Spain

ARTICLE INFO

Keywords:

Stress
Classification
EEG
Signal smoothing
Data leakage

ABSTRACT

Recently, multi-level stress assessment has become an active research subject. In this context, researchers typically develop models based on machine learning classifiers and features extracted from biosignals like electrocardiogram (ECG) or electroencephalogram (EEG). For that purpose, EEG power spectral density (PSD) is a recurrent feature owing to its high responsiveness and remarkable performance. However, PSD is usually smoothed to cope with its bursty nature, what may cause data leakage and hence call into question classification performance. In this study, our aim was twofold: first, to examine the effect of EEG-PSD smoothing in three-level stress classification, and second, to evaluate the practical viability of a two-level stress detector without smoothing. To this end, we conducted participants through a stress-relax session while recording their EEG. Then, we estimated the EEG-PSD and used the stress reported by the participants as labels for classification. Initially, we developed a three-level stress classifier and examined the effect of smoothing on its performance. We found that classification performance was directly proportional to smoothing intensity (F1-score 0.61–0.94), and also that when smoothing was not applied to features, classification performance was insufficient for practical applicability (AUC < 0.7). We link this behavior to train-test contamination due to smoothing. Subsequently, we attempted two-level stress classification without smoothing. In this case, performance met the criteria for practical applicability (AUC = 0.76). This suggests that performance enhancement in three-level stress classification was caused by data leakage produced by smoothing, and hence, to render realistic stress classifiers each epoch should be processed individually.

1. Introduction

Currently, mental stress is an unavoidable concern that affects people on a global scale. According to the American Psychological Association [1], main sources of stress include health care, climate change, and safety. Mental stress can be triggered by several aspects of daily life, such as work, routine, and restless periods, and is usually linked to psychophysiological symptoms like headaches or fatigue [2], although other important health issues may appear [3]. In terms of economic impact, annual expenses associated to work miss and stress-related health issues are estimated around USD 300 billion. These facts justify the growing interest in stress early detection and classification.

According to literature, stress can be detected through several physiological markers, like galvanic skin response (GSR) [4–6],

electroencephalography (EEG) [7–10], electrocardiography (ECG) [11–15], and cortisol [16]. With regard to EEG, brain activity is usually acquired and processed to obtain markers linked with stress, such as power in Theta, Alpha, and Beta bands, relative Gamma [17–19], coherence, and asymmetry. This technique has been widely applied in emotion recognition and mental illness screening studies [8,20–22]. In the case of ECG and GSR, stress level assessment is derived from heartbeat and skin conductance, respectively. In this context, cortisol is considered a gold standard for stress assessment, as the level of this hormone that is secreted into saliva increases when people are stressed. Among these techniques, EEG is the most generally cited approach for the evaluation of stress. This may be due to its high responsiveness and temporal resolution [19], what supports the implementation of real-time solutions [23].

* Corresponding author at: Research Centre for Information and Communications Technologies, University of Granada, Spain.

E-mail address: malg@ugr.es (M.A. Lopez-Gordo).

<https://doi.org/10.1016/j.bspc.2021.102881>

Received 16 December 2020; Received in revised form 25 March 2021; Accepted 16 June 2021

Available online 23 June 2021

1746-8094/© 2021 The Author(s).

Published by Elsevier Ltd.

This is an open access article under the CC BY-NC-ND license

(<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Current research trends dwell on stress assessment through the development of classification models that are built on biomarkers extracted from EEG, ECG, GSR, etc. With regard to EEG, classification is accomplished through the combination of brain activity features and classifiers such as support vector machines (SVM) [8,24,25], artificial neural networks (ANN) [26], random forest (RF) [27,28], etc. [29,30]. In this context, classification performance report is paramount, since practical stress classifiers require a certain level of efficiency. Although main approaches typically consider the classification of two levels of stress (stress and no stress), there are proposals for the discrimination of three or even more stress levels (typically low, medium, and high stress). For instance, in [31], authors aimed to evaluate the effect of music on the stress level of the participants from their EEG activity. They proposed a three-level stress classification considering low, medium, and high stress classes. The highest accuracy that they obtained for the three-level classification was around 95%. In the case of [24], the authors also achieved a remarkable performance for a five-level discrimination (the minimum accuracy obtained for one class was 90.26%). These examples support the potential of multi-level stress classification. Nevertheless, two important aspects are occasionally obviated when the results of a multi-level stress classifier are reported: a detailed explanation of the EEG processing and a meticulous report of the classifier performance. Usual EEG processing involves smoothing and filtering (e.g., Butterworth, Chebyshev, and Savitzky–Golay filters) what may produce data leakage when the data are splitted using common procedures such as grid search cross validation or k-fold cross validation. For instance, if multiple epochs are smoothed out together, information from adjacent epochs is shared. Thus, when most reproduced cross validation methods are applied, epochs in the train and test sets may contain correlated information, what leads to unrealistic good performances. Therefore, to render reliable and reproducible results, non-adequate signal processing must be avoided. Additionally, when classifier performance is reported, authors often provide only accuracy, hence disregarding the performance obtained in the discrimination of each stress level. To provide a comprehensive interpretation of the model performance, other metrics such as precision, recall, and F1 score should be considered.

In this paper, we examined the effect that smoothing filter window length has on the performance of a three-level stress classifier. We also evaluated a three-level and a two-level stress classifier from individually processed epochs and assessed their practical applicability. To this effect, we computed the EEG-PSD during a stress-relax session, and we performed two and three-level classifications of the stress perceived by a group of participants. We considered different performance metrics aside from accuracy, namely precision and recall for each class, and weighted F1 score. Finally, we discussed our results and examined other studies in literature.

2. Material and methods

2.1. Participants

Twenty healthy volunteers (14 males, 6 females, mean age 24.20 ± 4.03) participated in the study. They belonged to the community of the University of Granada (mostly students and staff members) and were recruited via email distribution lists. The participants did not suffer from any health condition neither mental disorder. They were asked to sign an informed consent, and to avoid any stimulant or relaxant the day before the experiment. Furthermore, participants were not rewarded in any way for their participation in the study. Each participant took part in a single study session that lasted roughly 18 min. The entire data capture was completed in approximately two weeks.

2.2. Experimental design

Once we briefed the participants about the different phases of the study, we equipped them with an EEG acquisition system to record their

EEG activity. The timeline of the experiment is illustrated in Fig. 1.

First, the participants completed a two-minute eyes-closed resting state period. Then, they performed the Montreal imaging Stress task (MIST), a test specifically designed to induce psychosocial stress [32] that has been widely validated in literature [33–35]. Subsequently, participants were randomly separated into two groups. The first group was conducted through a relaxation program that used a loop of three ambient light colors (blue, magenta, and green) inside a chromotherapy room. The second group experienced an immersive virtual reality app that simulated the chromotherapy room program. Finally, the participants completed another two-minute resting state period. Throughout the experiment, we required the participants to report their self-perceived stress level (SPSL) via surveys. These surveys are referred as T1-T4 in Fig. 1.

2.3. Experimental setup

Regarding the MIST, we implemented the test using MATLAB R2016a (MathWorks, USA) as a graphical interface controlled with the touchscreen of a laptop. To complete the test, the participants sat in a chair and were asked to use only their dominant hand. The MIST lasted nine minutes, including a three-minute training period and a six-minute test period. To record the EEG activity, we used the RABio w8 acquisition system, that works at a sampling rate of 500 Hz. Our team at the University of Granada developed this device, and we have successfully utilized it in previous works [23,36]. We located the electrodes at Fp1, Fp2, F7, F8, Fz, Cz, O1, and O2 positions of the 10–20 International System. Nevertheless, we only considered frontal and pre-frontal positions (Fp1, Fp2, F7, and F8) for further analysis in this study. We defined this electrode montage according to previous successful studies on stress assessment [17–19,23]. We referenced and grounded the electrodes to the left ear lobe.

In relation to the SPSL surveys, we conceived an adaptation of the perceived stress scale (PSS) to minimize the time required to answer. Only two questions were asked to the participants: “Is your current stress level higher or lower than the last time that we asked?” (possible answers “lower”, “higher”, and “equal”) and “What is your stress level in a scale from 1 to 5?” (possible answers 1 to 5, where 1 is the minimum level of stress and 5 is the maximum).

2.4. Signal processing

Upon completion of the data capture, we processed the EEG data of the participants offline. First, we filtered the EEG signals using a zero-phase shift 2nd order Butterworth filter with bandpass 1–50 Hz. Then, we applied a notch filter at 50 Hz to remove electric coupling. Thereafter, we splitted the EEG signals into two-second epochs without overlapping. To reject artifacts, we zeroed the epochs above a pre-arranged threshold of 100 μV . We selected this threshold following visual inspection and in compliance with previous EEG studies [37,38]. Subsequently, we detrended and z-scored each epoch, and estimated the channel-averaged PSD in five frequency bands (see Table 1) using the periodogram (no overlapping). For RS1 and RS2 (see Fig. 1), we processed only the central minute, hence the total time span of the processed signals was approximately 16 min.

In addition, we calculated the Alpha asymmetry as the difference between the Alpha power at prefrontal electrodes (Fp1 and Fp2), and the relative Gamma as indicated in Equation (1). Furthermore, according to previous studies on stress assessment by relative Gamma [17–19], we inverted the gamma power for participants S01, S04, S06, S09, S10, S12, S13, S15, S18, and S19 after visual inspection.

$$RG = \frac{P_{\text{Gamma}}}{P_{\text{Alpha}} + P_{\text{Theta}}} \quad (1)$$

Overall, we derived seven EEG-PSD signals for each participant. To equal the span of the signals from all the participants, we resampled

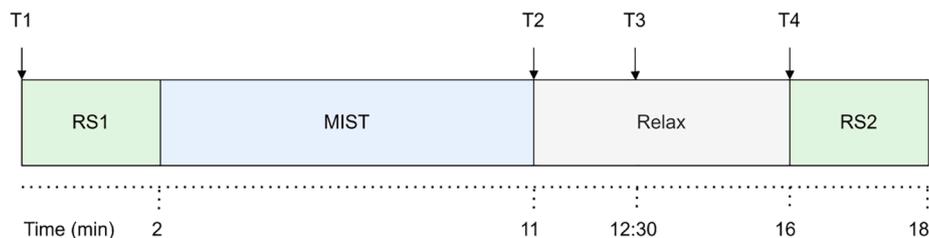


Fig. 1. Timeline of the experimental process. First the participants completed a two-minute resting state period (RS1). Then, they performed the Montreal imaging stress task (MIST). Subsequently, they went through a relax session (Relax). Finally, the participants completed another two-minute resting state period (RS2). Several surveys (T1-T4) were fulfilled by the participants to grade their self-perceived stress level (SPSL). The total span of the experiment was 18 min.

Table 1

Frequency bands where we obtained the PSD of the EEG.

Band	Range (Hz)
Delta	1–4
Theta	4–8
Alpha	8–13
Beta	13–25
Gamma	25–45

them to 480 samples (what corresponds to the aforementioned 16-minute span divided into two-second epochs). Lastly, we smoothed the seven signals using a 2nd order Savitzky-Golay filter. This procedure is widely applied in EEG literature to cope with the high temporal variability of brain electrical activity. For this filter, we examined window lengths of 11, 21, 31, 41, 51, and 61 epochs in order to assess the effect of this parameter on stress level classification performance. We also examined the case where each epoch was processed individually (referred as 0 window length throughout the paper). To illustrate the smoothing effect, Fig. 2 shows the application of a Savitzky-Golay filter with different window lengths to the EEG Delta power.

To conclude this subsection, it is worth to mention that EEG signals are affected by contamination sources such as MEG power and EMG power of forehead muscles. Decoupling of these signals would likely increase stress level classification performance. However, this procedure is out of the scope of this study, nonetheless the interested reader may refer to [32] and [33] for a comprehensive understanding about this topic.

2.5. Feature extraction

For each of the seven PSD signals extracted, we partitioned the

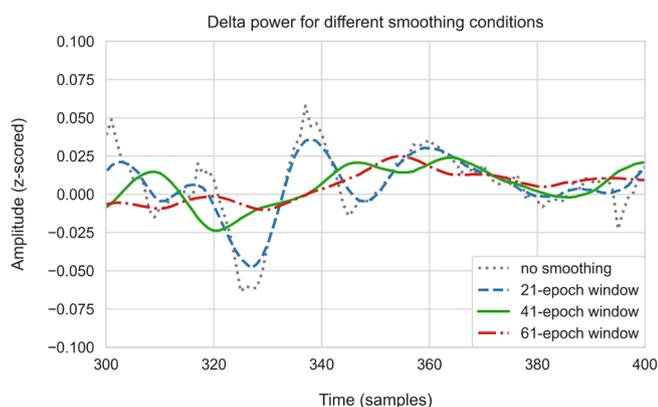


Fig. 2. Effect of smoothing Delta power using a 2nd order Savitzky-Golay filter. The curves in the graph correspond to smoothing windows of length 0, 21, 41, and 61 epochs (dotted gray, dashed blue, solid green, and dashed-dotted red, respectively).

minute preceding the SPSL surveys (T1-T4) into five segments, each one enclosing six epochs. Then, we averaged those six epochs for each segment, and we assigned it a label equal to the SPSL survey under consideration. Consequently, for each participant, we obtained a data matrix with 140 samples (4 surveys \times 5 segments \times 7 PSD signals), and one label per sample (values in the range 1–5 from the SPSL surveys). For classification, we combined the data matrices from all the participants. Fig. 3 represents the segmentation procedure described in this paragraph for one of the seven PSD signals.

2.6. Classification

As stated in the Introduction, in this study we examined a three-level classification task and a two-level stress detection. For the three-level classification task, we aimed to predict three stress levels (low stress, medium stress, and high stress). To this end, we merged the SPSL survey answers into three classes: low stress (labels 1 and 2), medium stress (labels 3 and 4), and high stress (label 5). For two-level stress detection, we intended to discern between two states (stress and no stress). Thus, in this case we combined the SPSL survey answers into two classes, namely, no stress (labels 1, 2, and 3) and stress (labels 4 and 5).

In regard to classification, first we splitted feature matrix X and target array y into stratified train and hold-out sets (80% train, 20% hold-out). We used the train set to find the best combination of hyperparameters for multiple classifiers, and the hold-out set to evaluate the classifier that achieved the best performance. Subsequently, we applied standardization and oversampling to the feature matrix. We implemented oversampling through SMOTE (Synthetic Minority Over-sampling Technique), as the high stress class (label 5) had a lower presence in the data compared to the rest. Essentially, this technique synthesizes new instances from the minority class by interpolating real instances of that class. After interpolation, we applied grid search cross-validation (GSCV) to find the combination of classifier and hyperparameters that best predicted SPSL from spectral features. The classifiers considered in this study included logistic regression (LR), support vector machine (SVM), random forest (RF), k-nearest neighbors (KNN), and multi-layer perceptron (MLP). We configured GSCV to perform a 5-fold cross-validation over all the possible combinations of hyperparameters for each classifier. For this procedure, we selected the weighted F1 score as scoring metric. In multiclass problems, this score is calculated per class and then averaged to obtain a general metric for the classifier performance. We selected the weighted approximation of the F1 score to account for class imbalance.

Lastly, as stated in subsection 2.4, we considered different window lengths for the Savitzky-Golay filter (0, 11, 21, 31, 41, 51, and 61 epochs) to assess the effect that this parameter had on three-level stress classification performance. To sum up, Fig. 4 represents the procedures described through this section.

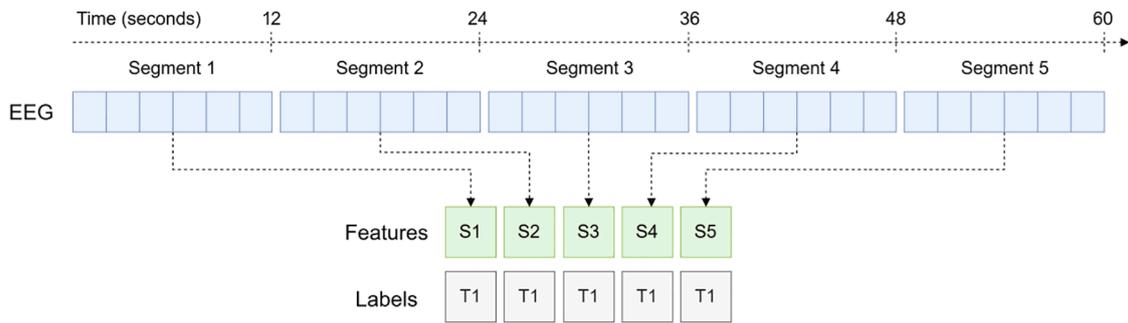


Fig. 3. PSD segmentation. We partitioned the minute preceding each SPSL survey into five segments each one holding six epochs (in blue). Then, we averaged each segment (in green) and we assigned it a label (in gray) corresponding to the SPSL of the corresponding survey. This figure is referred to the first survey that we conducted in the study (T1).

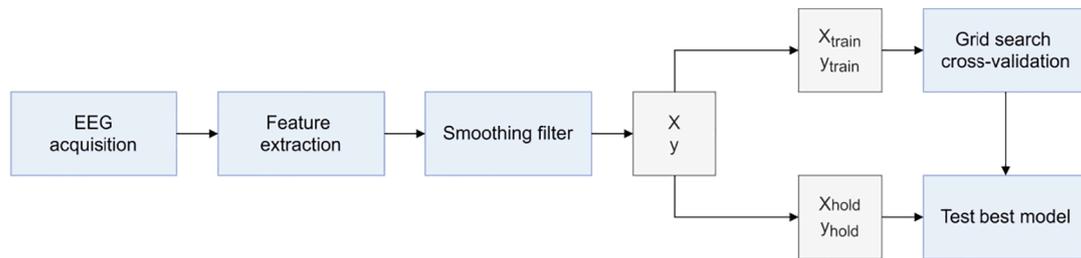


Fig. 4. Stress classification pipeline. First, we performed the EEG acquisition using RABio w8 system. Then, we estimated the PSD in different frequency bands, and we applied a Savitzky-Golay filter to smooth spectral data. Subsequently, we splitted the minute preceding the SPSL surveys (T1-T4) into five segments, and we estimated the average spectral features in each of them. Then, we reshaped the spectral features and survey answers to build feature matrix X and target array y. After that, we splitted these two data structures into training and hold-out sets. Thereafter, for each classifier analyzed in the study, we performed GSCV using only the training set. Finally, we assessed the best classifier on the hold-out set.

3. Results

3.1. Cross-validation

Table 2 presents GSCV results for the three-level stress classification task. In this table, we report the F1 score obtained by each of the five classifiers for each of the smoothing window lengths examined in this study.

Table 3 shows the GSCV results for the stress detection task. In this case, we did not apply a smoothing filter. The left column corresponds to the classifiers, and the right column displays the F1 score obtained by each of the models in the stress detection task.

Fig. 5 represents a graphic arrangement of Table 2, and is intended to facilitate its overview. It displays the F1 score obtained by the classifiers during GSCV for each of the smoothing window lengths considered in

Table 2

Results obtained during GSCV for the three-level stress classification task. The left most column represents the classifiers. The rest of the columns display the average F1 score (\pm std) obtained by each classifier for the different smoothing window lengths reported in epochs (0–61). Shadowed, the best classifier for each smoothing window length.

F1 score per smoothing window length							
Model	0	11	21	31	41	51	61
LR	0.45 ± 0.02	0.45 ± 0.03	0.48 ± 0.04	0.52 ± 0.02	0.55 ± 0.07	0.54 ± 0.05	0.56 ± 0.05
SVM	0.55 ± 0.04	0.63 ± 0.04	0.63 ± 0.05	0.72 ± 0.05	0.78 ± 0.07	0.84 ± 0.02	0.89 ± 0.05
RF	0.60 ± 0.05	0.62 ± 0.03	0.62 ± 0.08	0.67 ± 0.07	0.66 ± 0.08	0.74 ± 0.06	0.76 ± 0.05
KNN	0.54 ± 0.03	0.63 ± 0.05	0.63 ± 0.04	0.68 ± 0.03	0.74 ± 0.06	0.83 ± 0.04	0.82 ± 0.03
MLP	0.59 ± 0.02	0.61 ± 0.08	0.64 ± 0.06	0.68 ± 0.06	0.74 ± 0.08	0.75 ± 0.08	0.78 ± 0.03

Table 3

Results obtained during GSCV for the stress detection task. Left column indicates the models analyzed. Right column represents the average F1 score (\pm std) for each of the models analyzed. Shadowed, the classifiers that obtained the highest performance.

Model	F1 score
LR	0.65 ± 0.06
SVM	0.84 ± 0.02
RF	0.84 ± 0.03
KNN	0.78 ± 0.04
MLP	0.84 ± 0.02

the three-level stress classification task.

3.2. Test

Table 4 presents the performance of the best model yielded by GSCV on the hold-out set. The top part of the table corresponds to three-level classification and the second part refers to stress detection. The first column represents the different smoothing window lengths analyzed. For each of these lengths, the rest of the columns indicate the best classifier and its performance in terms of class precision and recall, F1 score, and accuracy.

Fig. 6 illustrates the receiver operating characteristic (ROC) curves of the best model for each smoothing window length considered in three-level stress classification. Each graph includes three curves, that represent the binary classification problem for each class versus the rest.

Fig. 7 depicts the ROC curve for the best model in the stress detection task.

Fig. 8 displays the F1 score obtained on the hold-out set by the best model for each of the smoothing window lengths analyzed in the three-

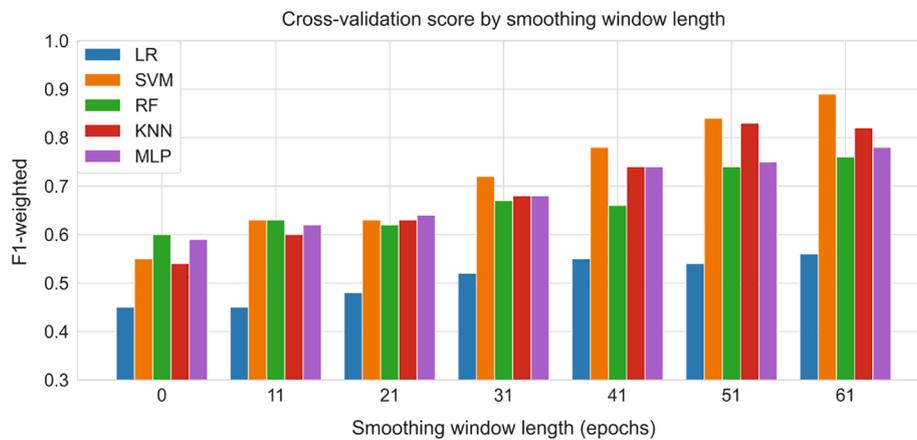


Fig. 5. GSCV score yielded by each classifier for the three-level stress classification task. Y axis corresponds to the F1 score, and X axis refers to the window length of the Savitzky-Golay filter used during EEG-PSD smoothing. Each bar in the graph represents a classifier: LR (blue), SVM (orange), RF (green), KNN (red), and MLP (purple).

Table 4

Performance metrics on the hold-out set for three-level classification (top) and stress detection (bottom). Reported metrics include precision (P) and recall (R) for each class, F1 score, and accuracy.

Smoothing window length	Best model	Class 1 (P)	Class 1 (R)	Class 2 (P)	Class 2 (R)	Class 3 (P)	Class 3 (R)	F1 score	Accuracy
<i>Three-level stress classification</i>									
0	RF	0.68	0.70	0.58	0.46	0.29	0.67	0.61	0.61
11	SVM	0.71	0.81	0.72	0.54	0.25	0.33	0.68	0.69
21	MLP	0.74	0.76	0.57	0.54	0.67	0.67	0.67	0.67
31	SVM	0.84	0.86	0.77	0.71	0.75	1	0.81	0.81
41	SVM	0.89	0.86	0.78	0.75	0.6	1	0.83	0.83
51	SVM	0.92	0.95	0.91	0.83	0.75	1	0.91	0.91
61	SVM	0.95	0.95	0.92	0.92	1	1	0.94	0.94
<i>Stress detection</i>									
0	MLP	0.92	0.85	0.38	0.56	–	–	0.83	0.81

level stress classification. We included this figure to illustrate the enhancement in performance as the window length is increased.

4. Discussion

In this study we have assessed the effect of smoothing filter window length on the performance of a multi-level stress classifier. Moreover, we have examined the practical appropriateness of a three-level stress classifier and a stress detector when each EEG-PSD segment is processed individually. With respect to three-level stress classification, for the non-smoothed case (zero epochs), the best model yielded an accuracy of 0.61. Alternatively, for the longest smoothing window length (61 epochs), the accuracy obtained by the best model was 0.94. In regard to the stress detection task, the best performing model yielded an accuracy of 0.81. These results evidence that, (i) the smoothing of the EEG-PSD improves classification performance, possibly due to train-test contamination; (ii) although the performance of our three-level stress classifier with no smoothing is not good enough for practical implementations, stress can be successfully detected by means of a two-level classification from non-smoothed EEG-PSD signals.

4.1. Cross-validation results

According to the results presented in Table 2, with scarcely smoothed data (filter window lengths 0–31 epochs) all the classifiers performed similarly, with the exception of LR. However, for longer windows (lengths 31–61 epochs), SVM performed noticeably better than the rest of the classifiers. These findings suggest that SVM outperforms other EEG-based multi-level classifiers when intense smoothing is applied to EEG data. This is coherent with other results reported in the literature

for multi-level stress classification [8,24,25,27]. With respect to stress detection, SVM, RF, and MLP yielded the highest performance, while KNN reached reasonable performance and LR discriminated worst. For the stress detection task, we processed each epoch individually, what prevents from spreading correlated information between adjacent epochs, and enhances the timing capabilities of the system. For instance, for a 61-epoch window, classification requires a 1-minute delay as a result of the collection of the 30 epochs posterior to the central sample. Conversely, for the non-smoothed approach, we used segments of 6 consecutive epochs, and hence, classification can be performed with a latency of 6 s and 12-second time resolution.

4.2. Test results

Table 4 shows the results of the test phase. In literature, it is frequent to report just the accuracy of three-level stress classifiers [8,25]. Nevertheless, to offer a comprehensive performance report, other relevant metrics such as precision and recall are required. To this end, in Table 4 we have reported accuracy, F1 score, and intra-class precision and recall. For both the three-level classification and stress detection tasks, the performance of the best classifier for all smoothing window lengths yielded results similar to those obtained in the cross-validation phase (confront the F1 score in Tables 2 and 3 with the same metric in Table 4).

To gain insights about the prediction capacities of the classifiers, we computed the ROC curves. Fig. 6 shows the ROC curves for all the smoothing window lengths considered in the three-level stress classification task. If we consider the clinical environment criterion in regard to AUC [39]: acceptable (AUC 0.7–0.8); excellent (AUC 0.8–0.9), we could interpret the results of Table 4 and Fig. 6 as follows:

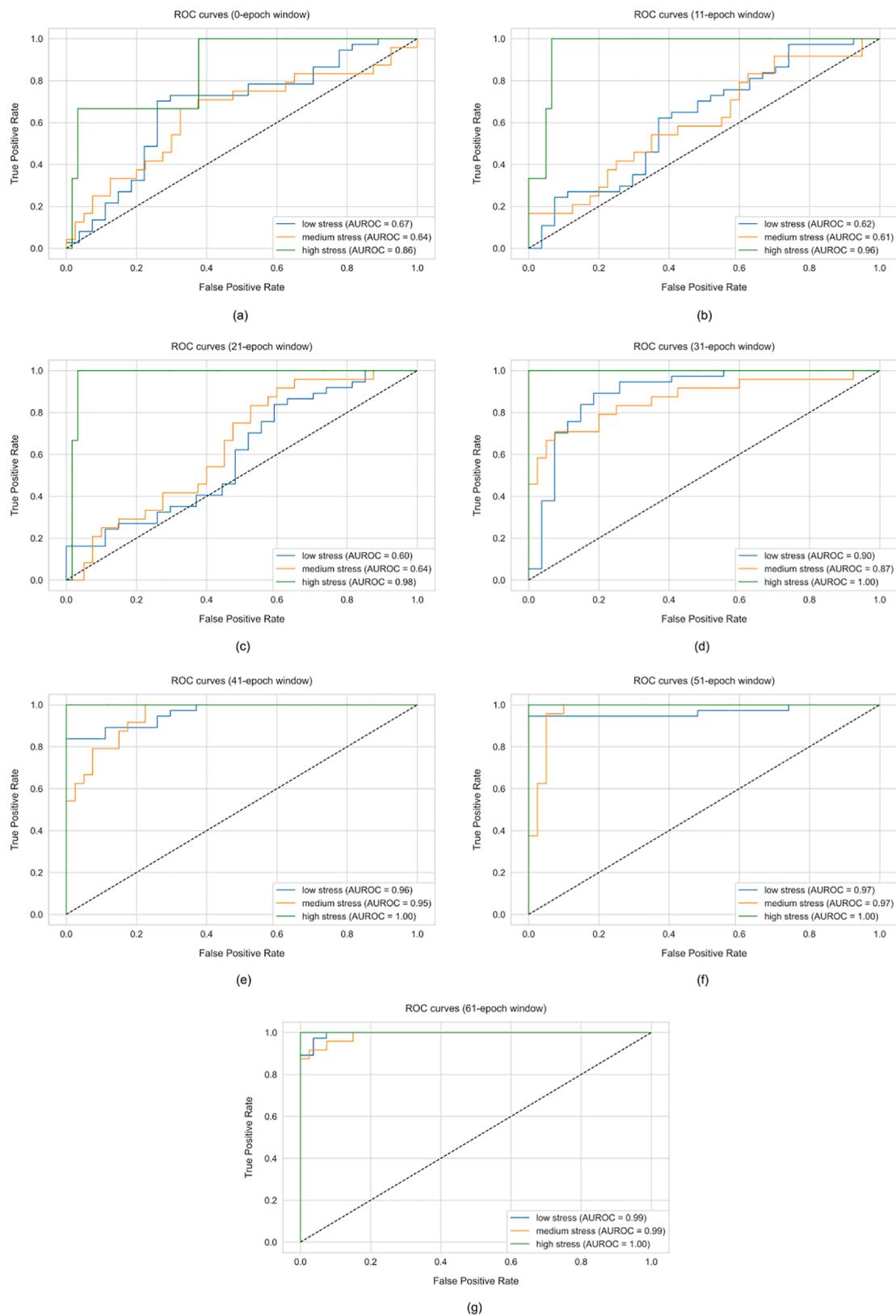


Fig. 6. ROC curves for three-level stress classification task. Smoothing window lengths of 0 (a), 11 (b), 21 (c), 31 (d), 41 (e), 51 (f), and 61 (g) epochs. Each graph includes three curves, representing the binary problem for a class versus the rest of the classes, for low stress (blue), medium stress (orange), and high stress (green). The area under the curve (AUC) is reported in brackets. Black dashed line represents the chance level.

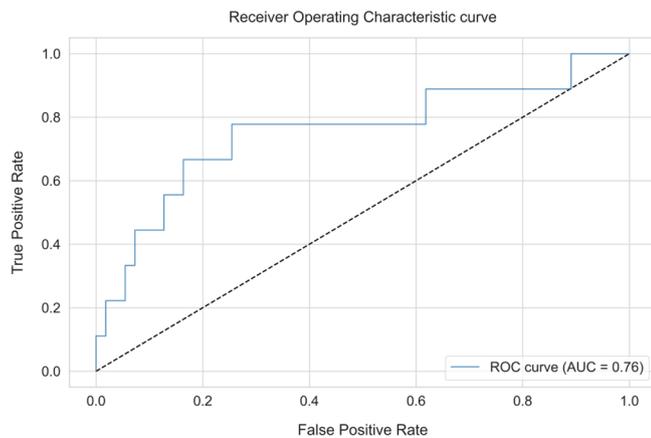


Fig. 7. ROC curve for the binary classification problem (stress vs. no stress). AUC is reported in brackets. Black dashed line represents the chance level.

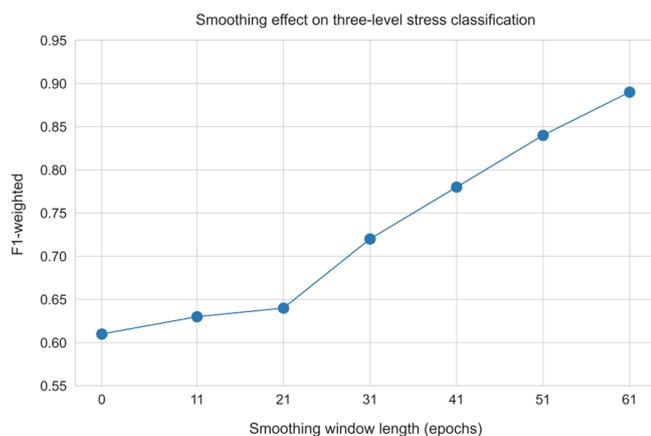


Fig. 8. Smoothing effect on three-level stress classification performance. X axis refers to the length of the smoothing filter window in epochs. Y axis represents the weighted F1 score yielded on the hold-out set by the best model.

- Three-level stress classification: a smoothing window length of 31 epochs is enough to achieve an excellent prediction capability of the three classes (low-medium-high AUC 0.90–0.87–1.00, respectively). With lengths shorter than 31 epochs, the quality of our classifier is below acceptable for low and medium stress classes (AUC between 0.60 and 0.64). Regardless the smoothing window length, the detection of high stress level is always excellent (AUC between 0.86 and 1).
- Stress detection: the detection of stress from non-smoothed data can be described as acceptable (AUC 0.76).

In view of the results, we can conclude that our methodological approach is reproducible and the stress detection we proposed in this study constitutes a feasible alternative for practical implementations of a two-level stress classifier. With regard to three-level stress classification, our non-smoothed classification approach is close to acceptable for low and medium levels of stress (AUC 0.67 and 0.64, respectively), and excellent for high level of stress (AUC 0.86). The reason behind the remarkable performance in the discrimination of the high stress class might be justified by more energetic EEG-PSD features than the other stress levels, although this assumption requires further analysis. Consequently, our non-smoothed three-level classification approach is not appropriate for practical implementations of multi-level stress classifiers. For wider smoothing windows (31–61 epochs), the performance of the three-level stress classifiers improves considerably. However, as stated before, when smoothing is applied, the classification may

incur in train-test contamination, as the information of adjacent EEG segments is shared, and diminish timing capabilities.

4.3. Comparative discussion

In literature, many studies have reported high performance for multi-level stress classification. However, a direct comparison with our results is not trivial. In some of those studies, authors only reported accuracy [8,25] and did not provide an analysis of the intra class performance (e.g., confusion matrix, precision, recall, or ROC curves). This is not an appropriate way to report the performance of a three-level stress classifier. For instance, a classifier with a very high sensibility and low specificity for a high-probability class, and null sensibility and specificity for a low-probability class, could obtain a very high accuracy. In addition, although segmentation and filtering greatly influence classification performance, these aspects of the EEG processing are often overlooked. In [31], the authors declared that they applied a 75% overlapping method (unspecified) to the EEG signals to obtain the PSD, and reached an accuracy of 98.76%. Other studies reported accuracies up to 93.6% and 94.3% [24,25], but the authors only indicated that the EEG spectral bands were denoised without providing additional details. Figs. 5 and 8 demonstrate that three-level stress classification performance increases as the smoothing window is enlarged. Therefore, a comparative discussion with other studies that do not accurately report the EEG processing is inherently unviable.

5. Conclusions

In this work, we evaluated the performance of a three-level stress classifier for different smoothing filter window lengths, and we also approached three-level and two-level stress classification from non-smoothed EEG-PSD data. To this effect, we recorded the brain activity of participants during a stress-relax session and we extracted EEG-PSD features. We used self-perceived stress level surveys as labels for classification. We assessed multiple stress classifiers whose hyperparameters were derived via GSCV. To evaluate the effect of smoothing on three-level stress classification, we considered smoothing filter window lengths ranging from 0 to 61 epochs, and we estimated multiple performance metrics (accuracy, F1 score, and intra-class precision and recall). The results we obtained do not support the appropriateness of a three-level stress classifier from non-smoothed EEG data for practical applications. Our results also evidence that data smoothing increases classification performance at the expense of train-test contamination and poorer timing capabilities. According to our results, it is feasible to classify three stress levels from individually processed EEG segments with remarkable performance (AUC 0.67, 0.64, and 0.86 for low, medium, and high stress classes). In terms of stress detection, we implemented a two-level stress classifier without smoothing of adjacent segments that achieved close to excellent performance (AUC 0.76). This supports the implementation of our two-level stress approach in practical stress detection solutions. In this regard, an EEG stress classifier could be extended to real-time applications, avoiding the need to enquire the participants about their SPSL. These applications may include stress relief therapies or neuromarketing, as in these environments stress is considered an important concern, and there is still a lack of stress anticipation solutions. Lastly, we provide the following recommendations for future EEG-based stress classification studies: (i) performance of three and two-level stress classifiers could be further enhanced if the EEG spectral features were combined with other features, such as galvanic skin response or heart rate variability; (ii) each EEG segment should be processed individually, so the content of adjacent segments is not combined; (iii) an appropriate classifier performance report should include metrics aside from accuracy, like intra-class precision and recall, F1 score, or AUC. In the future, reliable three-level stress classifiers could be used in brand new scenarios such as offices, schools, or even at home.

CRediT authorship contribution statement

Eduardo Perez-Valero: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing - original draft, Writing - review & editing, Visualization. **Miguel A. Lopez-Gordo:** Conceptualization, Methodology, Validation, Resources, Writing - review & editing, Supervision, Project administration. **Miguel A. Vaquero-Blasco:** Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the project PGC2018-098813-B-C31 (the Spanish Ministry of Science, Innovation and Universities, by European Regional Development Funds and by the Nicolo Association for the R&D in Neurotechnology for disability). The authors would also like to thank Dr. Maria Jose Sanchez Carrion from the School for Special Education San Rafael of Granada, Hospitaller Order of St. John of God, for their support and for providing access to their chromotherapy room. We also want to thank Prof. Alexander Bertrand, from the Department of Electrical Engineering (ESAT), Stadius Center for Dynamical Systems, Signal Processing and Data Analytics, KU Leuven, (Belgium) for his suggestions and valuable comments for this manuscript.

References

- [1] Stress in America™ 2019, p. 9.
- [2] N.B. Anderson et al., American Psychological Association, p. 23.
- [3] S. Cohen, et al., Chronic stress, glucocorticoid receptor resistance, inflammation, and disease risk, *Proc. Natl. Acad. Sci. U.S.A.* 109 (16) (2012) 5995–5999, <https://doi.org/10.1073/pnas.1118355109>.
- [4] A. Fernandes R. Helawar R. Lokesh T. Tari A.V. Shahapurkar Determination of stress using Blood Pressure and Galvanic Skin Response 2014 Sivakasi, India, Dec 165 168 10.1109/CNT.2014.7062747.
- [5] H. Kurniawan, A.V. Maslov, M. Pechenizkiy, Stress detection from speech and Galvanic Skin Response signals, in: Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems, Porto, Portugal, Jun. 2013, pp. 209–214, doi: 10.1109/CBMS.2013.6627790.
- [6] M.V. Villarejo, B.G. Zapirain, A.M. Zorrilla, A stress sensor based on galvanic skin response (GSR) controlled by ZigBee, *Sensors* 12 (5) (May 2012) 6075–6101, <https://doi.org/10.3390/s120506075>.
- [7] H. Díaz M., F. M. Cid, J. Otárola, R. Rojas, O. Alarcón, L. Cañete, EEG Beta band frequency domain evaluation for assessing stress and anxiety in resting, eyes closed, basal conditions, *Procedia Comp. Sci.*, 162 (2019), 974–981, doi: 10.1016/j.procs.2019.12.075.
- [8] H. Jebelli, S. Hwang, S. Lee, EEG-based workers' stress recognition at construction sites, *Autom. Constr.* 93 (2018) 315–324, <https://doi.org/10.1016/j.autcon.2018.05.027>.
- [9] A.C. Marshall, N.R. Cooper, The association between high levels of cumulative life stress and aberrant resting state EEG dynamics in old age, *Biol. Psychol.* 127 (2017) 64–73, <https://doi.org/10.1016/j.biopsycho.2017.05.005>.
- [10] J. Minguiillon, M.A. Lopez-Gordo, D.A. Renedo-Criado, M.J. Sanchez-Carrion, F. Pelayo, Blue lighting accelerates post-stress relaxation: results of a preliminary study, *PLoS ONE* 12 (10) (2017), e0186399, <https://doi.org/10.1371/journal.pone.0186399>.
- [11] L. Chen, Y. Zhao, P. Ye, J. Zhang, J. Zou, Detecting driving stress in physiological signals based on multimodal feature analysis and kernel classifiers, *Expert Syst. Appl.* 85 (2017) 279–291, <https://doi.org/10.1016/j.eswa.2017.01.040>.
- [12] P. Karthikeyan, M. Murugappan, S. Yaacob, Detection of human stress using short-term ecg and hrv signals, *J. Mech. Med. Biol.* 13 (02) (2013) 1350038, <https://doi.org/10.1142/S0219519413500383>.
- [13] N. Keshan, P.V. Parimi, I. Bichindaritz, Machine learning for stress detection from ECG signals in automobile drivers, in: 2015 IEEE International Conference on Big Data (Big Data), Oct, 2015, pp. 2661–2669, <https://doi.org/10.1109/BigData.2015.7364066>.
- [14] S. Pourmohammadi, A. Maleki, Stress detection using ECG and EMG signals: a comprehensive study, *Comput. Methods Programs Biomed.* 193 (2020), 105482, <https://doi.org/10.1016/j.cmpb.2020.105482>.
- [15] M.N. Rastgoo, B. Nakisa, F. Maire, A. Rakotonirainy, V. Chandran, Automatic driver stress level classification using multimodal deep learning, *Expert Syst. Appl.* 138 (2019), 112793, <https://doi.org/10.1016/j.eswa.2019.07.010>.
- [16] M. Annerstedt, et al., Inducing physiological stress recovery with sounds of nature in a virtual reality forest — results from a pilot study, *Physiol. Behav.* 118 (2013) 240–250, <https://doi.org/10.1016/j.physbeh.2013.05.023>.
- [17] A. Lutz, L.L. Greischar, N.B. Rawlings, M. Ricard, R.J. Davidson, Long-term meditators self-induce high-amplitude gamma synchrony during mental practice, *Proc. Natl. Acad. Sci.* 101 (46) (2004) 16369–16373, <https://doi.org/10.1073/pnas.0407401101>.
- [18] S.R. Steinhubl, et al., Cardiovascular and nervous system changes during meditation, *Front. Hum. Neurosci.* 9 (2015), <https://doi.org/10.3389/fnhum.2015.00145>.
- [19] J. Minguiillon, M.A. Lopez-Gordo, F. Pelayo, Stress assessment by prefrontal relative gamma, *Front. Comput. Neurosci.* 10 (2016), <https://doi.org/10.3389/fncom.2016.00101>.
- [20] A.R. Aslam, T. Iqbal, M. Aftab, W. Saadeh, M.A.B. Altaf, A10.13uJ/classification 2-channel deep neural network-based SoC for emotion detection of autistic children, in: 2020 IEEE Custom Integrated Circuits Conference (CICC), Mar, 2020, pp. 1–4, <https://doi.org/10.1109/CICC48029.2020.9075952>.
- [21] A.R. Aslam, M.A.B. Altaf, An on-chip processor for chronic neurological disorders assistance using negative affectivity classification, *IEEE Trans. Biomed. Circuits Syst.* 14 (4) (2020) 838–851, <https://doi.org/10.1109/TBCAS.2020.3008766>.
- [22] A.R. Aslam, M.A.B. Altaf, An 8 channel patient specific neuromorphic processor for the early screening of autistic children through emotion detection, in: 2019 IEEE International Symposium on Circuits and Systems (ISCAS), May, 2019, pp. 1–5, <https://doi.org/10.1109/ISCAS.2019.8702738>.
- [23] J. Minguiillon, E. Perez, M. Lopez-Gordo, F. Pelayo, M. Sanchez-Carrion, Portable system for real-time detection of stress level, *Sensors* 18 (8) (2018) 2504, <https://doi.org/10.3390/s18082504>.
- [24] M. Zubair, C. Yoon, Multilevel mental stress detection using ultra-short pulse rate variability series, *Biomed. Signal Process. Control* 57 (2020), 101736, <https://doi.org/10.1016/j.bspc.2019.101736>.
- [25] S. Lotfan, S. Shadyad, R. Khosrowabadi, A. Mohammadi, B. Hatfe, Support vector machine classification of brain states exposed to social stress test using EEG-based brain network measures, *Biocybern. Biomed. Eng.* 39 (1) (2019) 199–213, <https://doi.org/10.1016/j.bbe.2018.10.008>.
- [26] H. Sharma, R. Raj, M. Juneja, EEG signal based classification before and after combined Yoga and Sudarshan Kriya, *Neurosci. Lett.* 707 (2019), 134300, <https://doi.org/10.1016/j.neulet.2019.134300>.
- [27] Z. Halim, M. Rehan, On identification of driving-induced stress using electroencephalogram signals: a framework based on wearable safety-critical scheme and machine learning, *Information Fusion* 53 (2020) 66–79, <https://doi.org/10.1016/j.inffus.2019.06.006>.
- [28] Y.-W. Kim, et al., Riemannian classifier enhances the accuracy of machine-learning-based diagnosis of PTSD using resting EEG, *Prog. Neuro-Psychopharmacol. Biol. Psychiatry* 102 (2020), 109960, <https://doi.org/10.1016/j.pnpbp.2020.109960>.
- [29] O. Attalah, An effective mental stress state detection and evaluation system using minimum number of frontal brain electrodes, *Diagnostics* 10 (5) (2020) 292, <https://doi.org/10.3390/diagnostics10050292>.
- [30] M. Shim, M.J. Jin, C.-H. Im, S.-H. Lee, Machine-learning-based classification between post-traumatic stress disorder and major depressive disorder using P300 features, *NeuroImage: Clinical* 24 (2019), 102001, <https://doi.org/10.1016/j.nicl.2019.102001>.
- [31] A. Asif, M. Majid, S.M. Anwar, Human stress classification using EEG signals in response to music tracks, *Comput. Biol. Med.* 107 (2019) 182–196, <https://doi.org/10.1016/j.compbiomed.2019.02.015>.
- [32] K. Dedovic, R. Renwick, N. K. Mahani, V. Engert, S. J. Lupien, J.C. Pruessner, The Montreal Imaging Stress Task: using functional imaging to investigate the effects of perceiving and processing psychosocial stress in the human brain, *J. Psychiatry Neurosci.*, p. 7.
- [33] A. Brugnera, C. Zarbo, M.P. Tarvainen, P. Marchettini, R. Adorni, A. Compare, Heart rate variability during acute psychosocial stress: a randomized cross-over trial of verbal and non-verbal laboratory stressors, *Int. J. Psychophysiol.* 127 (2018) 17–25, <https://doi.org/10.1016/j.ijpsycho.2018.02.016>.
- [34] L. Han, Q. Zhang, X. Chen, Q. Zhan, T. Yang, Z. Zhao, Detecting work-related stress with a wearable device, *Comput. Ind.* 90 (2017) 42–49, <https://doi.org/10.1016/j.compind.2017.05.004>.
- [35] A.R. Subhani, W. Mumtaz, M.N.B.M. Saad, N. Kamel, A.S. Malik, Machine learning framework for the detection of mental stress at multiple levels, *IEEE Access* 5 (2017) 13545–13556, <https://doi.org/10.1109/ACCESS.2017.2723622>.
- [36] M.A. Vaquero-Blasco, E. Perez-Valero, M.A. Lopez-Gordo, C. Morillas, Virtual reality as a portable alternative to chromotherapy rooms for stress relief: a preliminary study, *Sensors*, 20 (2020) 21, doi: 10.3390/s20216211.
- [37] C. Jeunet et al., Uncovering EEG correlates of covert attention in soccer goalkeepers: towards innovative sport training procedures, *Sci. Rep.*, 10 (1) (2020), Art. no. 1, doi: 10.1038/s41598-020-58533-2.
- [38] R. Cassani, T.H. Falk, F.J. Fraga, M. Cecchi, D.K. Moore, R. Anghinah, Towards automated electroencephalography-based Alzheimer's disease diagnosis using portable low-density devices, *Biomed. Signal Process. Control* 33 (2017) 261–271, <https://doi.org/10.1016/j.bspc.2016.12.009>.
- [39] J.N. Mandrekar, Receiver operating characteristic curve in diagnostic test assessment, *J. Thoracic Oncol.* 5 (9) (2010) 1315–1316, <https://doi.org/10.1097/JTO.0b013e3181ec173d>.