Feature Article:
DOI. No. 10.1109/MAES.2021.3115198

Human Stress Detection With Wearable Sensors Using Convolutional Neural Networks

Manuel Gil-Martin, Ruben San-Segundo, Ana Mateos, and Javier Ferreiros-Lopez, **Ciudad Universitaria, 28040 Madrid, Spain**

INTRODUCTION

Stress can be considered as a mental state that provokes physical or mental tension. In the aviation industry, there are many professionals with a high level of stress at work like pilots or air traffic controllers. An excess of stress level can eventually focus his or her attention too much, reducing the awareness of other aspects. This situation can increase the risk of human errors [1]. Additionally, This factor has been associated to the development of several illness like obesity [2], diabetes [3], asthma [4], or cardiovascular diseases [5].

Stress produces physiological responses generating characteristic patterns in some biosignals like skin conductance, heart rate, or body temperature [6], [7]. Stress detection can be done by using signal processing and machine learning techniques to identify characteristic patterns in these biosignals. Automatic stress monitoring has important benefits. First, automatic detection can help people to manage stress situations, reducing the effect on their work, health, or daily activities [8], like driving [9], [10]. Moreover, physicians could have objective metrics to evaluate the stress exposure of patients or to develop intelligent medical applications that adapt their behavior to the level of stress (i.e., blood glucose predictor for diabetic patients).

The fast development of wearable technologies and the wide expansion of smart devices have facilitated the recording of biosignals during daily activities [11]. These technologies have permitted the design and implementation of low-cost stress supervision systems based on wearable sensors. This article contributes proposing and evaluating a stress detection system based on deep learning using wearable sensors. The main contributions can be detailed in the following points.

- First, we propose and evaluate a deep learning architecture based on convolutional neural networks (CNNs) for stress monitoring. This architecture has a first part that includes three convolutional layers for extracting features from inertial and physiological signals. The second part is composed of three fully connected layers for stress detection.
- Second, we analyze several signal processing techniques to be applied before defining the inputs to the deep learning architecture: Fourier transforms, cube root, and constant Q transform (CQT).

These contributions were evaluated on a public dataset, wearable stress and affect detection (WESAD) [12], recorded by Robert Bosch GmbH Corporate Research in Germany. This evaluation - validation. This analysis allows estimating the system performance when monitoring subjects who are not considered in the training set. This article analyses the different contributions in three classification tasks with two classes (stress versus nonstress), three classes (stress versus baseline versus amusement), and five classes (stress versus baseline versus amusement versus meditation versus recovery). These classification tasks were also addressed in previous works [12], [13]. This article reports the best results on these tasks using a LOSO cross-validation.

This article is organized as follows. The "Related Work on Stress Detection Using Biosignals" section includes related work on stress detection. "Material and Methods" describes the material and methods used in this study: the dataset and signal processing and deep learning modules. The experiments and the obtained results are detailed in "Results." Finally, the "Conclusion" section summarizes the main conclusions of this article.

Authors' current address: Manuel Gil-Martin, Ruben San-Segundo, Ana Mateos, and Javier Ferreiros-Lopez, E.T.S.I. Telecomunicacion, Ciudad Universitaria, 28040 Madrid, Spain (e-mail: manuel.gilmartin@upm.es, ruben.sansegundo@upm.es, ana.mateossa@alumnos.upm.es, javier.ferreiros@upm.es).
Manuscript received September 28, 2020, revised July 6, 2021; accepted September 6, 2021, and ready for publication September 22, 2021.
Review handled by Stefan Brueggenwirth.

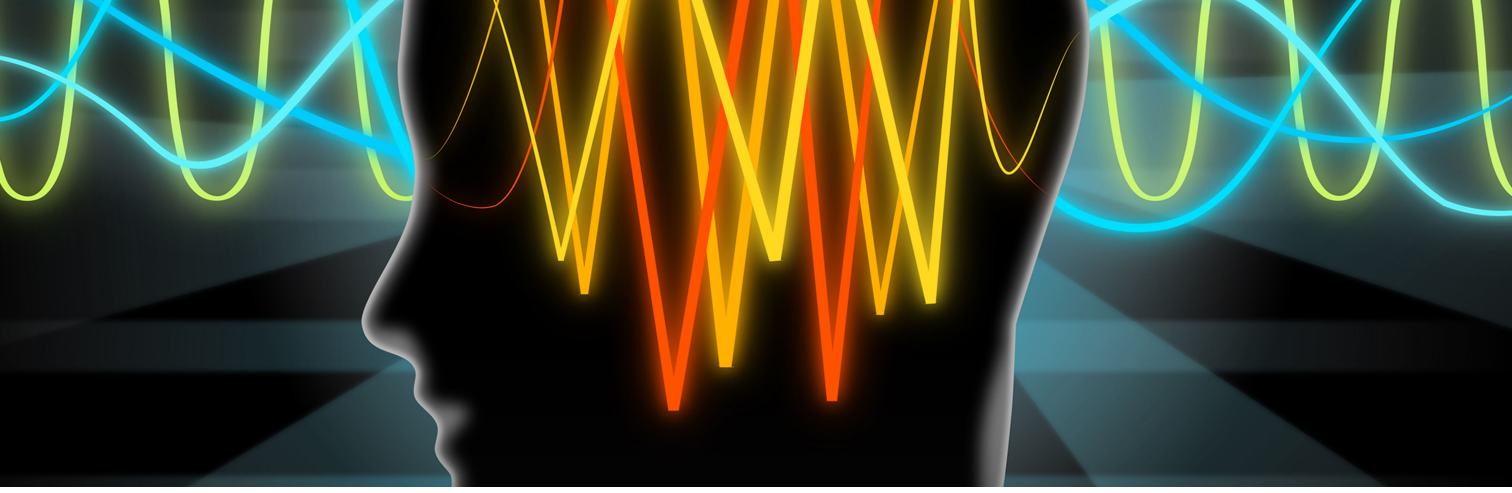

Image licensed by Ingram Publishing

RELATED WORK ON STRESS DETECTION USING BIOSIGNALS

This section reviews related work in stress detection using biosignals. Many previous works focused on extracting handcrafted features from physiological signals like electroencephalograph [14], electrocardiogram (ECG), electromyogram (EMG), and electrodermal activity (EDA), and physical signals like respiration, temperature, or eye activity. The review provided by Giannakakis *et al.* [15] describes a summary of the main characteristics obtained from these biosignals.

Features obtained from ECGs include metrics obtained from QRS curves, heart rate, heart rate variability, or energy distribution in the frequency domain [16]–[22]. ECG and EDA are the biosignals most frequently used for stress monitoring. EDA has two main components: skin conductance level (SCL) and skin conductance response (SCR). SCL shows the slow variation and SCR the short-term reaction to some stimuli. Main EDA features are extracted from SCR curves [23]–[26]: number of peaks, peak duration, peak value, slope, etc. Regarding EMG most of the features are extracted from the energy distribution in a wide range of frequencies reaching several hundreds of Hz [27].

These physiological signals can be complemented with physical signals like respiration [16], [29], [28] or body temperature [21]. In both cases, the main features are obtained from simple statistics along a certain time. In the case of respiration, the relation between inspiration and expiration time can also provide stress information [29].

Thanks to the fast expanding of wearable devices, new datasets including a high number of biosignals have become available to the research community: SWELL [30], Multimodal Stress Detection [31], WESAD [12] in an academic environment [32]. When combining several biosignals for stress detection, machine learning algorithms can help to deal with a more complex task of finding stress patterns in a multimodal scenario. In the literature, there are several previous works that analyzed and compared several machine learning algorithms [12], [13], [17], [25], [32], including Support Vector Machine, Random Forest, Decision Trees, Regression Strategies, Boost algorithms, K-Nearest Neighbors, etc. The best results were obtained with the combination of several of these algorithms. Some references include interesting comparison tables including relevant characteristics of each work [26], [32]: signals, features, and machine learning algorithms and performance.

The application of deep learning algorithms is increasing very fast due to the better performance reached with these techniques. Deep neural networks can be used to detect stress patterns, but also, they can be used to learn and combine relevant features from several biosignals. In the literature, there are studies on stress detection that used deep learning algorithms for classifying handcrafted features [13], [19]. Other studies used recurrent neural networks to learn features and classify stress patterns from ECGs [16] or ECG in combination with car driving signals [34]. In this work, we propose a deep learning architecture based on CNNs to learn relevant features from a high number of biosignals (14 biosignals) and detect stress patterns. We also study several alternatives for processing the signals before being considered as inputs to the CNNs.

Regarding the system architecture for pervasive healthcare applications, previous works have focused on analyzing the evolution of information and communication technology for the improvement of quality of life [35]. In fact, some works have identified the most representative healthcare scenarios that can benefit from 5G networks [36] and they have proposed cascaded networks to handle healthcare traffic in assisted-living applications for chronic patients [37]. The architecture of this work is based on computing different signal processing techniques for each type of signal and feed a CNN with multimodal sensed data. Afterward, this CNN architecture combines this information and learn dependencies among the different signals to perform the classification task.

The proposed architecture was evaluated on the WESAD dataset using a LOSO cross-validation, similar to Schmidt *et al.*'s and Chakraborty *et al.*'s previous works [12], [13]. Schmidt *et al.* [12] used handcrafted features for each biosignal, evaluating several machine learning algorithms. They reported results on two classification tasks: stress versus nonstress and baseline versus stress versus amusement.

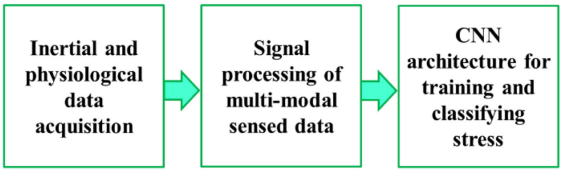

Figure 1.
General architecture of the stress detection system.

Chakraborty *et al.* [13] used a deep neural network to detect stress patterns from handcrafted features. They evaluated their proposal on a classification task considering five classes: baseline, stress, amusement, meditation, and recovery. The results presented in this article improved the results reported in these two previous works on this dataset.

MATERIAL AND METHODS

The stress detection system developed in this work consists of three main modules: the first one collects the inertial and physiological signals from different devices, the second one processes these multimodal data to define the best inputs to the deep neural network, and the third module includes the deep neural network that extracts the main features and detect stress periods. Figure 1 shows the general architecture of the stress detection system. In this section, we comment on the dataset, the signal processing methods analyzed in this article, and the deep learning structure considered for stress detection.

DATASET DESCRIPTION

In this study, we used a public dataset, WESAD [12]. This dataset includes recordings from 15 subjects (12 males and 3 females) wearing two wearable devices: a RespiBAN Professional and Empatica E4. The RespiBAN device, on the chest, recorded body acceleration (in three axes), body temperature, respiration, ECG, EMG, and EDA. All these signals were sampled at 700 Hz. The Empatica E4 bracelet (on the wrist) measured hand acceleration (in three axes), arm temperature, blood volume pulse (BVP), and EDA. These signals were recorded with different sampling frequencies, but all of them were upsampled to 64 Hz using the Fourier method. The evaluation subjects were selected excluding people with mental or cardiovascular problems, pregnant people, or heavy smokers. The mean age was 27.5.

After synchronizing and instating all the devices in the subject's body, all subjects followed a protocol in a laboratory considering several phases:

- Baseline phase: the subject was asked to remain in a neutral position at a table (sitting or standing) reading some magazines for 20 minutes.
- Amusement phase: the participant watched funny videos with a duration of 6 minutes.
- Stress phase: the stress phase was provoked by exposing the subjects to a trier social stress test [38] for 10 minutes. This test consisted of a public speaking and a mental arithmetic task. First, each subject was asked to speak about their personal strengths and weaknesses in front of human resource people, and second, each participant was asked to subtract 17 from 2023 to zero without error.
- Meditation phase: the subjects were guided by an expert to perform breathing exercises with closed eyes while sitting in a comfortable position to reestablish them to neural affective mood.
- Recovery phase: both devices were synchronized again and removed from the subject's body.

Regarding the specific postures of the subjects while the data collection protocol, the baseline, amusement, and stress phases were conducted either standing or sitting (half of the subjects were standing and half were sitting for each phase). During the meditation phase, all subjects were sitting.

Considering previous works on stress monitoring [12], [13] using this dataset, three different classification tasks have been considered. These are the classification tasks analyzed in this work:

- In the first classification task [12], only data from three phases were considered: baseline, stress, and amusement phases. The target was to detect stress (stress phase) from nonstress (baseline and amusement phases) (S versus NS).
- The second classification task [12] consisted of distinguishing between baseline, stress, and amusement phases (B versus S versus A).
- Finally, the third task [13] focused on discriminating between 5 classes: baseline, stress, amusement, meditation, and recovery phases (B versus S versus A versus M versus R).

SIGNAL PROCESSING

All signals from wearable devices (RespiBAN on the chest and Empatica E4 on the wrist) were segmented in 60-second windows with a shift of 0.25 seconds. The classification is performed at the window level. Previous works [12], [13] used the same windowing, extracting handcrafted features from each window to be passed to the classifier. In this work, we did not extract handcrafted features, but we processed the different signals before passing them to the CNN. The convolutional layers in the CNN were responsible for learning appropriate features for stress detection.

For all signals, we subdivided each 60-second window into P-second subwindows with a shift of 0.25 seconds. We computed the Fourier transform (using the Fast

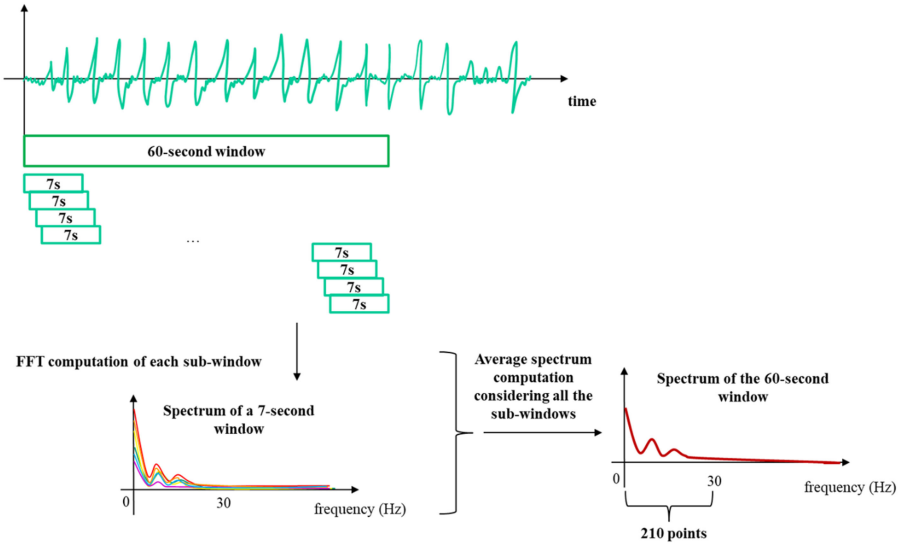

Figure 2.
Signal processing steps for RespiBAN accelerometer signals.

Fourier Transform, FFT) and obtained the spectrum (module of the Fourier transform) of each subwindow, and then we averaged the spectrum along all subwindows in a 60-second window. Finally, we selected a frequency range of the average spectrum and the spectrum coefficients included in this range were the inputs to the CNN.

We computed different processing techniques depending on each signal to accommodate the heterogeneous signals into a vector with the same length and feed the CNN. These signals had different sampling frequencies and contained relevant information within different frequency ranges. For these reasons, each 60-second window was subdivided into P-second subwindows and the spectrum was computed from each subwindow. To represent the whole 60-second window, the average spectrum was computed along the subwindows. The subwindow length depends on the signal and was adjusted (according to the signal sampling rate) with the target of generating an average spectrum with the same number of frequency points: for all signals, the number of frequency points M was equal to 210. This number of points was obtained by multiplying the subwindow length by the upper bound of the frequency range of each signal. This computation allowed stacking all inputs in a N x M matrix, where N is the number of signals and M is the number of frequency points representing each 60-second window (210 in our case). For example, for the RespiBAN accelerometer, each 60-second window of three accelerations (in *X*-, *Y*-, and *Z*-axes) were subdivided in 7-second subwindows shifted 0.25 seconds. The frequency range in these cases was 0–30 Hz, the motion information above 30 Hz is negligible. For 7-second subwindows, we obtained a spectrum with a resolution of seven points per Hz. Considering a range of 0–30 Hz, we obtained 210 spectral points as inputs to the CNN. Figure 2 shows the signal processing steps for acceleration signals, where a 60-second window is subdivided into 7-second subwindows and the spectrum is computed for each sub-window within the range 0–30 Hz. Afterward, the average spectrum is computed along the subwindows to represent a 60-second window with 210 frequency points. Table 1 summarizes all the details per signal. In all cases, the number of inputs to the CNN is the same, 210.

The first variant we considered was to calculate the cube root (CR) of subwindow spectrums before averaging in a 60-second window. The cube root allows emphasizing harmonics with low energy remarking the spectral information.

The second variant consisted in computing CQT [39] in subwindows. This CQT was implemented as a postprocessing step after Fourier transform, as suggested by Brown *et al.* [40]. CQT transforms a time-domain signal into a frequency representation where the frequency bins are geometrically spaced with equal Q-factors. The Q factor is the ratio of the center frequency and the bandwidth. CQT provides more resolution at low frequencies and defines the same distance between consecutive harmonics, independently of the fundamental frequency. Maintaining the same distance between consecutive harmonics facilitates the learning process of convolutional filters in the CNN. In this work, we have considered 21 bins for postprocessing the 210 frequency points obtained after Fourier transform and frequency range selection. F represents the frequency filters used to obtain the 21 CQT bins. We applied the same frequency filters to postprocess the 210 points of all signals. As the selected 210 points corresponded to different frequency ranges, the CQT is automatically adapted to the frequency range of each signal. After CQT application, the

Table 1.

Processing Details Per Signal					
Device	Signal	Sampling frequency	Frequency range	Sub-window length	Number of inputs to CNN
RespiBAN (Chest)	Accelerations (*X*, *Y*, and *Z*)	700 Hz	0–30 Hz	7 seconds	210
	ECG	700 Hz	0–7 Hz	30 seconds	210
	EDA	700 Hz	0–7 Hz	30 seconds	210
	EMG	700 Hz	0–250 Hz	0.84 seconds	210
	RESP	700 Hz	0–6 Hz	35 seconds	210
	TEMP	700 Hz	0–6 Hz	35 seconds	210
Empatica E4 (Wrist)	Accelerations (X, Y and Z)	64 Hz	0–30 Hz	7 seconds	210
	BVP	64 Hz	0–7 Hz	30 seconds	210
	EDA	64 Hz	0–7 Hz	30 seconds	210
	TEMP	64 Hz	0–6 Hz	35 seconds	210

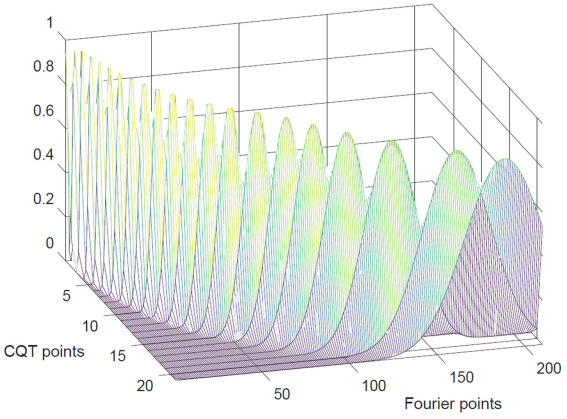

Figure 3.
CQT filters for postprocessing Fourier transform.

number of inputs to the CNN per signal was reduced from 210 to 21. Figure 3 displays the filters that relate the Fourier points and the CQT points.

CONVOLUTIONAL NEURAL NETWORKS

Figure 4 represents the CNN used in this work. Table 2 summarizes the settings for all layers. The first part of the CNN learns relevant features from signal spectra. This part is composed of three convolutional layers with two intermediate maxpooling layers. With the target of not increasing a lot the number of parameters, we considered kernel sizes of (1 x 3) in all convolutional layers. For simplicity, the padding characteristic was defined as "same": the input and output shapes are the same.

The classification part includes three fully connected layers for classification, with a decreasing number of units. To avoid overfitting, we included dropout layers (with a fraction of 0.3) after convolutional and fully connected layers. ReLU is the activation function in intermediate layers reducing the impact of gradient vanishing effect.

The CNN input is a 2 D matrix N x M, where N is the number of signals and M the number of points per signal. N varies depending on the number of signals considered in the experiment and M is different if we consider Fourier transform (210 points) of the CQT (21 points). The number of outputs corresponds to the number of classes considered in the classification problem, using the softmax function and the categorical cross-entropy loss metric.

We used the training set (using part of the training set for validation) to tune the number of epochs (10) and the batch_size (50). The optimizer was fixed with the root-mean-square propagation method [41].

RESULTS

The experimental setup is the same than in previous works [12], [13], we used a LOSO cross validation along the 15 subjects: 14 subjects were used for training and 1 for testing. This process was repeated 15 times, leaving a different subject for testing every time.

For evaluating the results, we considered accuracy and F1-score in percentages. The accuracy is the ratio between the number of correctly classified examples and the number of total samples. The F1-score is the harmonic mean of precision and recall. The final F1-score is the average

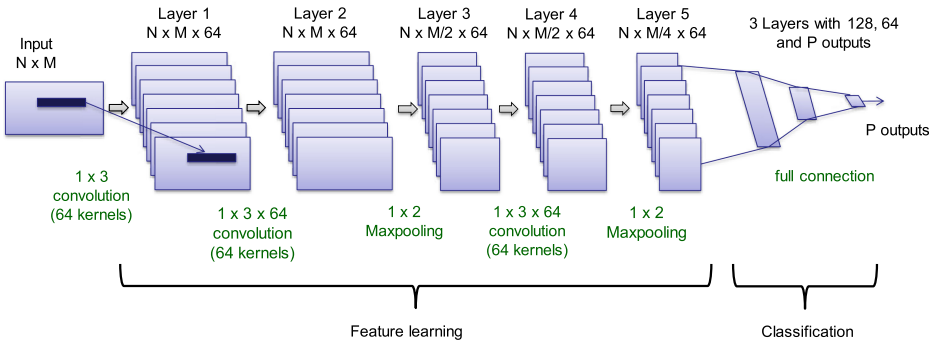

Figure 4.
Deep learning structure including convolutional and fully connected layers.

Table 2.

Configuration Details and Number of Parameters to Train for all Layers Considering an Input with Shape N = 14 x M = 210				
Layer	Output shape	Param #	Activation function	Other characteristics
Input	(-, 14, 210, 1)	–	–	–
Feature extraction				
2D Conv	(-, 14, 210, 64)	256	ReLU	# Kernels = 64, Size = 1 x 3, Stride = 1, Padding = 'same'
Dropout	(-, 14, 210, 64)	–	–	portion = 0.3
2D Conv	(-, 14, 210, 64)	12352	ReLU	# Kernels = 64, Size = 1 x 3 x 64, Stride = 1, Padding = 'same'
2D Max Pooling	(-, 14, 105, 64)	–	–	Pool size = 1 x 2
Dropout	(-, 14, 105, 64)	–	–	portion = 0.3
2D Conv	(-, 1, 105, 64)	12352	ReLU	# Kernels = 64, Size = 1 x 3 x 64, Stride = 1, Padding = 'same'
Dropout	(-, 1, 105, 64)	–	–	portion = 0.3
2D Max Pooling	(-, 1, 52, 64)	–	–	Pool size = 1 x 2
Dropout	(-, 1, 52, 64)	–	–	portion = 0.3
Classification				
Flatten	(-, 46592)	–	–	
Dense	(-, 128)	5963904	ReLU	# Neurons = 128, Initializer = 'glorot_uniform'
Dropout	(-, 128)	–	–	portion = 0.3
Dense	(-, 64)	8256	ReLU	# Neurons = 64, Initializer = 'glorot_uniform'
Dropout	(-, 64)	–	–	portion = 0.3
Dense	(-, P)	65 x P	Softmax	# Neurons = P, Initializer = 'glorot_uniform'
Output	(-, P)	–	–	–

Table 3.

Accuracy (%) and F1-score (%) Depending on the Signal Processing Module Considering all Signals in the Three Classification Tasks						
	S versus NS		B versus S versus A		B versus S versus A versus M versus R	
Signal processing	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
Fourier Transform	95.20 ± 0.13	95.13 ± 0.13	83.12 ± 0.13	82.67 ± 0.13	79.67 ± 0.25	79.24 ± 0.25
Fourier + Cube Root	96.73 ± 0.11	96.65 ± 0.11	85.00 ± 0.11	84.92 ± 0.22	81.21 ± 0.24	81.45 ± 0.24
Fourier + Cube Root + CQT	96.62 ± 0.11	96.63 ± 0.11	85.03 ± 0.22	85.01 ± 0.22	81.15 ± 0.24	81.70 ± 0.24

along the classes. All analyses were performed in the three classification tasks addressed in this paper (see "Dataset Description" section). Both evaluation metrics include confidence intervals (of 95%), obtained with the next equation [42]. ACC is the accuracy in percentage and N the number of examples (60-second windows) used in the evaluation. Two experiments are significantly different when there is not overlap in the confidence intervals computed with the following equation. N is the number of examples used in the test and ACC is the performing metric (Accuracy or F1-score) in percentage

$$CI\ \ (95\%) = \ \ \pm 1.96\sqrt{\frac{ACC\ \cdot (100 - ACC)}{N}}. \qquad (1)$$

The main parameters of the deep learning structure were adjusted subdividing the training set into training (13 subjects) and validation (1 subject) sets. Thirteen subjects were used for weights training while the validation subject was used for tuning the hyperparameters of the CNN. Once the CNN structure was defined, the weights were retrained using the 14 subjects.

SIGNAL PROCESSING ANALYSIS

Table 3 compares the performance obtained for the three classification tasks depending on the signal processing strategy. These results were obtained using all signals recorded from RaspiBAN and Empatica E4 devices. The first conclusion is that applying the cube root over the Fourier spectrum significantly improved the results. As commented before, the cube root allowed emphasizing frequencies with low energy remarking the harmonic structure. The CQT did not provide any additional improvement, but we obtained the same performance (high overlap in the confidence intervals) reducing significantly (10 times) the input shape: from 14 x 210 without CQT to 14 x 21 with CQT. To reduce the CNN input shape, we apply CQT in the next experiments.

COMPARISON WITH PREVIOUS STUDIES

In this section, we compare the performance of the proposed system with previous results reported in the literature. This comparison is done for the three classification tasks addressed in this work and detailing the results for different groups of signals. We grouped the signals recorded on the chest (RaspiBAN device) into inertial signals (acceleration in three axes) and physiological signals (ECG, EDA, EMG, temperature, and respiration). In a similar way, signals from the wrist (Empatica E4) were divided into inertial signals (accelerations in the three axes) and physiological signals (BVP, EDA, and temperature).

Table 4 compares our results with those obtained by Schmidt *et al.* [12] for the 2-class classification problem (stress versus nonstress), considering different groups of signals. Our results were significantly better than those obtained by Schmidt *et al.* for all groups of signals. In both studies, physiological signals from the chest were the best group of signals, while the inertial signals from the wrist was the worst. It is important to note how the CNN can integrate information from different devices and types of signals obtaining an important improvement when using all signals.

Table 5 shows the confusion matrix for the case of using all signals to classify between stress and nonstress windows. The accuracy is very high for both classes.

Table 6 shows the same comparison considering the 3-class classification task (baseline, stress, and amusement). Our results were significantly better than those obtained by Schmidt *et al.* [12] for all groups of signals. In this case, the physiological signals provided the best results, like those reported when combining all signals. Acceleration signals did not provide any help in this task.

Table 4.

Accuracy (%) and F1-Score (%) Comparison with Schmidt et al. [12] Classifying Between Stress and Non-Stress Situations

Stress versus Nonstress				
	Schmidt *et al.* [12]		This article	
	Accuracy	F1-score	Accuracy	F1-score
Chest inertial signals (acc)	73.87	62.12	90.60 ± 0.18	90.50 ± 0.18
Chest physiological signals	**93.12**	**91.47**	93.20 ± 0.16	92.70 ± 0.16
Wrist inertial signals (acc)	71.69	61.70	88.70 ± 0.20	88.50 ± 0.20
Wrist physiological signals	88.33	86.10	87.30 ± 0.21	87.00 ± 0.21
All inertial signals (acc)	–	–	91.50 ± 0.17	91.30 ± 0.17
All physiological signals	92.51	90.93	95.01 ± 0.13	94.78 ± 0.13
All chest signals	92.83	91.07	93.10 ± 0.16	93.01 ± 0.16
All wrist signals	87.12	84.11	92.70 ± 0.16	92.55 ± 0.16
All signals	92.28	90.74	**96.62 ± 0.11**	**96.63 ± 0.11**

Table 7 shows the confusion matrix for the case of using all signals to classify between baseline, stress, and amusement. The accuracy is very high when detecting stress, but this accuracy decreases very much for amusement, showing an important confusion between amusement and baseline.

Table 8 shows the results for the 5-class classification task and compares them with those reported by Chakraborty *et al.* [13]. Our results demonstrate the better performance of the system proposed in this article. Similar to the previous analysis, the physiological signals provided the best results, obtaining the same results as when combining all signals. Accelerations did not provide any additional improvement. In Chakraborty *et al.* [13], the authors used a deep learning structure, but the main difference was in the CNN input. While Chakraborty *et al.* extracted handcrafted features to feed the CNN, we used enhanced spectra as CNN inputs, leaving the CNN to learn specific features from those spectra.

Table 9 shows the confusion matrix for the case of using all signals to classify between baseline, stress, amusement, meditation, and recovery. The accuracy is very high when detecting stress. This accuracy is small for

Table 5.

Confusion Matrix Considering all Signals for Classifying Between Stress and Nonstress Situations

		Prediction	
		Nonstress	Stress
Ground Truth	**Nonstress**	96.5%	3.5%
	Stress	3.2%	96.8%

Table 6.

Accuracy (%) and F1-Score (%) Comparison with Schmidt et al. [12] Classifying Between Baseline, Stress, and Amusement Situations

Baseline versus Stress versus Amusement				
	Schmidt *et al.* [12]		This article	
	Accuracy	F1-score	Accuracy	F1-score
Chest inertial signals (acc)	56.56	44.28	68.32 ± 0.29	69.86 ± 0.29
Chest physiological signals	**80.34**	**72.51**	81.87 ± 0.24	81.21 ± 0.24
Wrist inertial signals (acc)	57.20	46.38	71.82 ± 0.28	71.23 ± 0.28
Wrist physiological signals	76.17	66.33	75.10 ± 0.27	74.00 ± 0.27
All inertial signals (acc)	–	–	72.23 ± 0.28	72.90 ± 0.28
All physiological signals	79.86	71.10	**85.10 ± 0.22**	**85.05 ± 0.22**
All chest signals	76.50	72.49	77.11 ± 0.26	76.90 ± 0.26
All wrist signals	75.21	64.12	74.90 ± 0.27	74.60 ± 0.27
All signals	79.57	68.85	**85.03 ± 0.22**	**85.01 ± 0.22**

Table 7.

Confusion Matrix Considering all Signals for Classifying Between Baseline, Stress and Amusement Situations

		Prediction		
		Baseline	Stress	Amusement
Ground Truth	**Baseline**	85.0%	2.5%	12.5%
	Stress	0.7%	98.3%	1.0%
	Amusement	25.8%	8.0%	66.2%

Table 8.

Accuracy (%) and F1-Score (%) Comparison with Chakraborty et al. [13] Classifying Between Baseline, Stress, Amusement, Meditation, and Recovery Situations

Baseline versus Stress versus Amusement versus Meditation versus Recovery				
	Chakraborty *et al.* [13]		This article	
	Accuracy	F1-score	Accuracy	F1-score
Chest inertial signals (acc)	–	–	54.91 ± 0.31	53.71 ± 0.31
Chest physiological signals	–	–	76.67 ± 0.26	75.23 ± 0.26
Wrist inertial signals (acc)	–	–	58.05 ± 0.31	57.12 ± 0.31
Wrist physiological signals	–	–	61.23 ± 0.30	62.34 ± 0.30
All inertial signals (acc)	–	–	59.78 ± 0.30	60.37 ± 0.30
All physiological signals	–	–	**82.12 ± 0.24**	**81.64 ± 0.24**
All chest signals	–	–	77.21 ± 0.26	77.67 ± 0.26
All wrist signals	–	–	70.02 ± 0.28	70.23 ± 0.28
All signals	**77.06**	**78.24**	**81.15 ± 0.24**	**81.70 ± 0.24**

amusement and recovery, showing an important confusion between baseline, amusement, and recovery.

ANALYSIS OF THE WINDOW LENGTH

All results reported in the previous sections were obtained using a 60-second window. This characteristic allowed the comparison with previous works on this dataset. To extend the analysis, we studied the influence of the window length in the three classification tasks (see Figure 5). In these experiments, all signals from both devices were considered to detect stress.

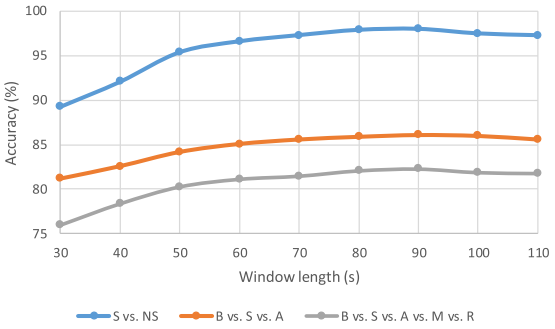

Figure 5.
Accuracy evolution depending on the window length for the three classification tasks. All signals were considered for classification.

The performance obtained with 60-second windows was improved (over 1%), increasing the analysis window to 90 seconds. On the one hand, we reached saturation in accuracy with longer windows. On the other hand, reducing to half the window length, the performance decreased significantly.

CONCLUSION

Automatic stress detection from biosignals is essential to evaluate stress situations and report objective metrics to physicians. These metrics can be very interesting to evaluate the stress exposure of the patients between consecutive visits. The main contributions of this article are the proposal of a deep learning architecture based on CNNs for stress detection, and the analysis of several signal processing techniques to generate the inputs to the deep learning architecture: Fourier

Table 9.

Confusion Matrix Considering all Signals for Classifying Between Baseline, Stress, Amusement, Meditation, and Recovery Situations

		Prediction				
		Baseline	Stress	Amusement	Meditation	Recovery
Ground Truth	**Baseline**	78.0%	1.9%	13.5%	2.5%	4.1%
	Stress	0.8%	96.7%	1.0%	1.0%	0.5%
	Amusement	23.3%	1.4%	48.8%	16.5%	10.0%
	Meditation	2.5%	0.5%	5.0%	90.5%	1.5%
	Recovery	13.0%	1.0%	19.0%	8.5%	58.5%

transform, cube root, and CQT. The CNN can learn features and detect stress patterns from several biosignals in parallel.

Regarding the signal processing techniques, the main conclusion is that the cube root over the Fourier spectrum significantly improved the results. The cube root emphasized the frequencies with low energy, making the harmonic structure more apparent. Although the CQT did not provide any additional improvement, we were to reduce 10 times the input shape to the CNN.

These analyses were performed on a public dataset, WESAD dataset, using a LOSO cross validation. We evaluated the contributions in three different classification tasks. Comparing the obtained results with previous works [12], [13], the accuracy increased from 93.1% to 96.6% classifying stress versus nonstress states. For the task of differencing between baseline versus stress versus amusement, the accuracy increased from 80.3% to 85.1%. In this classification task, the highest confusion appeared between nonstress states (baseline and amusement). Finally, comparing our results with those presented by Chakraborty *et al.* [13], the accuracy increased from 77.1% to 82.1% when dealing with five classes: baseline, stress, amusement, meditation, and recovery. The highest accuracy is obtained when detecting stress and the lowest when discriminating between baseline, amusement, and recovery. As future work, it would be very interesting to consider physiological devices located on other parts of the body to avoid disturbing the aviation professionals while monitoring their stress in practical applications.

CONFLICT OF INTEREST

Authors have no conflict of interest to declare.

ACKNOWLEDGMENTS

The work leading to these results is part of the projects AMIC (TIN2017-85854-C4-4-R) and CAVIAR (TEC2017-84593-C2-1-R) funded by MINECO/AEI/10.13039/501100011033 and by "ERDF A way of making Europe." The authors also thank Mark Hallett for the English revision of this paper and all the other members of the Speech Technology Group for the continuous and fruitful discussion on these topics. They gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used for this research.

REFERENCES

[1] P. Aricò *et al.*, "Human factors and neurophysiological metrics in air traffic control: A critical review," *IEEE Rev. Biomed. Eng.*, vol. 10, pp. 250–263, 2017.

[2] A. M. Heraclides, T. Chandola, D. R. Witte, and E. J. Brunner, "Work stress, obesity and the risk of type 2 diabetes: Gender-specific bidirectional effect in the whitehall II study," *Obesity*, vol. 20, no. 2, pp. 428–433, 2012.

[3] A. Mitra, "Diabetes and stress: A review," *Ethnomedicine*, vol. 2, no. 2, pp. 131–135, 2008.

[4] Y.-M. Oh, Y. S. Kim, S. H. Yoo, S. K. Kim, and D. S. Kim, "Association between stress and asthma symptoms: A population-based study," *Respirology*, vol. 9, no. 3, pp. 363–368, 2004.

[5] A. Steptoe and M. Kivimäki, "Stress and cardiovascular disease: An update on current knowledge," *Annu. Rev. Public Health*, vol. 34, no. 1, pp. 337–354, 2013.

[6] S. D. Kreibig, "Autonomic nervous system activity in emotion: A review," *Biol. Psychol.*, vol. 84, no. 3, pp. 394–421, 2010.

[7] R. W. Levenson and A. M. Ruef, "Physiological aspects of emotional knowledge and rapport," in *Empathic Accuracy*, W. J. Ickes, ed. New York, NY, USA: Guilford Press, 1997, pp. 44–72.

[8] L. P. S. Dias, J.,L. V. Barbosa, L. P. Feijó, and H. D. Vianna, "Development and testing of iAware model for ubiquitous care of patients with symptoms of stress, anxiety and depression," *Comput. Methods Programs Biomed.*, vol. 187, 2020, Art. no. 105113.

[9] L.-L. Chen, Y. Zhao, P.-F. Ye, J. Zhang, and J.-Z. Zou, "Detecting driving stress in physiological signals based on multimodal feature analysis and kernel classifiers," *Expert Syst. Appl.*, vol. 85, pp. 279–291, 2017.

[10] W. Hadi, N. El-Khalili, M. AlNashashibi, and G. Issa, "Abed Alkarim Albanna, application of data mining algorithms for improving stress prediction of automobile drivers: A case study in Jordan," *Comput. Biol. Med.*, vol. 114, 2019, Art. no. 103474.

[11] R. San-Segundo, H. Blunck, J. Moreno-Pimentel, A. Stisen, and M. Gil-Martín, "Robust human activity recognition using smartwatches and smartphones," *Eng. Appl. Artif. Intell.*, vol. 72, pp. 190–202, Jun. 2018.

[12] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. V. Laerhoven, "Introducing WESAD, a multimodal dataset for wearable stress and affect detection," in *Proc. Int. Conf. Multimodal Interact.*, Oct. 16–20, 2018, p. 9.

[13] S. Chakraborty, S. Aich, M.-I. Joo, M. Sain, and H.-C. Kim, "A multichannel convolutional neural network architecture for the detection of the state of mind using physiological signals from wearable devices," *J. Healthcare Eng.*, vol. 2019, 2019, Art. no. 5397814.

[14] A. Asif M. Majid, and S. M. Anwar, "Human stress classification using EEG signals in response to music tracks," *Comput. Biol. Med.*, vol. 107, pp. 182–196, 2019.

[15] G. Giannakakis, D. Grigoriadis, K. Giannakaki, O. Simantiraki, A. Roniotis, and M. Tsiknakis, "Review on psychological stress detection using biosignals," *IEEE Trans. Affect. Comput.*, early access, Jul. 9, 2019, doi: 10.1109/TAFFC.2019.2927337.

[16] W. Seo, N. Kim, S. Kim, C. Lee, and S.-M. Park, "Deep ECG-Respiration network (DeepER net) for recognizing mental stress," *Sensors*, vol. 19, 2019, Art. no. 3021.

[17] N. Keshan, P. V. Parimi, and I. Bichindaritz, "Machine learning for stress detection from ECG signals in automobile drivers," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, 2015, pp. 2661–2669.

[18] E. Jovanov, A. O'Donnell Lords, D. Raskovic, P. G. Cox, R. Adhami, and F. Andrasik, "Stress monitoring using a distributed wireless intelligent sensor system," *IEEE Eng. Med. Biol. Mag.*, vol. 22, no. 3, pp. 49–55, May/Jun. 2003.

[19] A. Oskooei, S. M. Chau, J. Weiss, A. Sridhar, M. R. Martínez, and B. Michel, "DeStress: Deep learning for unsupervised identification of mental stress in firefighters from Heart-rate variability (HRV) data," 2019. Arxivabs/1911.13213.

[20] G. Boateng and D. Kotz, "StressAware: An app for real-time stress monitoring on the amulet wearable platform," in *Proc. IEEE MIT Undergraduate Res. Technol. Conf.*, 2016, pp. 1–4.

[21] S. Barua, S. Begum, and M. U. Ahmed, "Supervised machine learning algorithms to diagnose stress for vehicle drivers based on physiological sensor signals," *Stud. Health Technol. Informat.*, vol. 211, pp. 241–248, 2015.

[22] A. H. Al-Jebrni, B. Chwyl, X. Y. Wang, A. Wong, and B. J. Saab, "AI-enabled remote and objective quantification of stress at scale," *Biomed. Signal Process. Control*, vol. 59, 2020, Art. no. 101929.

[23] J. Healey and R. Picard, "Detecting stress during real-world driving tasks using physiological sensors," *IEEE Trans. Intell. Transp. Syst.*, vol. 6, no. 2, pp. 156–166, Jun. 2005.

[24] J. Choi, B. Ahmed, and R. Gutierrez-Osuna, "Development and evaluation of an ambulatory stress monitor based on wearable sensors," *IEEE Trans. Inf. Technol. Biomed.*, vol. 16, no. 2, pp. 279–286, Mar. 2012.

[25] M. V. Villarejo, B. G. Zapirain, and A. M. Zorrilla, "A stress sensor based on galvanic skin response (GSR) controlled by zigbee," *Sensors (Basel)*, vol. 12, no. 5, pp. 6075–6101, 2012.

[26] K. Kyriakou *et al.*, "Detecting moments of stress from measurements of wearable physiological sensors," *Sensors (Basel)*, vol. 19, no. 17, 2019, Art. no. 3805.

[27] J. Wijsman, B. Grundlehner, and H. Hermens, "Trapezius muscle EMG a spredictor of mental stress," in *Proc. Wireless Health*, 2010, pp. 155–163.

[28] J. R. M. Fernández and L. Anishchenko, "Mental stress detection using bioradar respiratory signals," *Biomed. Signal Process. Control*, vol. 43, pp. 244–249, 2018.

[29] K. Plarre, A. Raij, and M. Scott, "Continuous inference of psychological stress from sensory measurements collected in the natural environment," in *Proc. 10th Int. Conf. Inf. Process. Sensor Netw.*, 2011, pp. 97–108.

[30] S. Koldijk, M. Sappelli, S. Verberne, M. Neerincx, and W. Kraaij, "The SWELL knowledge work dataset for stress and user modeling research," in *Proc. 16th ACM Int. Conf. Multimodal Interaction*, Nov. 12-16, 2014, pp. 291–298.

[31] K. Pisanski *et al.*, "Multimodal stress detection: testing for covariation in vocal, hormonal and physiological responses to Trier Social Stress Test," *Hormones Behav.*, vol. 106, pp. 52–61, 2018.

[32] M. Gjoreski, M. Luštrek, M. Gams, and H. Gjoreski, "Monitoring stress with a wrist device using context," *J. Biomed. Informat.*, vol. 73, pp. 159–170, 2017.

[33] J. Rodríguez-Arce, L. Lara-Flores, O. Portillo-Rodríguez, and R. Martínez-Méndez, "Towards an anxiety and stress recognition system for academic environments based on physiological features," *Comput. Methods Programs Biomed.*, vol. 190, 2020, Art. no. 105408.

[34] M. N. Rastgoo, B. Nakisa, F. Maire, A. Rakotonirainy, and V. Chandran, "Automatic driver stress level classification using multimodal deep learning," *Expert Syst. Appl.*, vol. 138, 2019, Art. no. 112793.

[35] G. Cisotto and S. Pupolin, "Evolution of ICT for the improvement of quality of life," *IEEE Aerosp. Electron. Syst. Mag.*, vol. 33, no. 5–6, pp. 6–12, May/Jun. 2018.

[36] G. Cisotto, E. Casarin, and S. Tomasin, "Requirements and enablers of advanced healthcare services over future cellular systems," *IEEE Commun. Mag.*, vol. 58, no. 3, pp. 76–81, Mar 2020, doi: 10.1109/mcom.001.1900349.

[37] S. Martiradonna, G. Cisotto, G. Boggia, G. Piro, L. Vangelista, and S. Tomasin, "Cascaded WLAN-FWA networking and computing architecture for pervasive in-home healthcare," 2020. ArXiv, vol. abs/2010.03805.

[38] C. Kirschbaum, K. Pirke, and D. Hellhammer, "The trier social stress test – a tool for investigating psychobiological stress responses in a laboratory setting," *Neuropsychobiology*, vol. 28, no. 1–2, pp. 76–81, 1993.

[39] J. C. Brown, "Calculation of a constant q spectral transform," *J. Acoust. Soc. Am.*, vol. 89, no. 1, pp. 425–434, 1991.

[40] J. C. Brown and M. S. Puckette, "An efficient algorithm for the calculation of a constant q transform," *J. Acoust. Soc. Am.*, vol. 92, no. 5, pp. 2698–2701, 1992.

[41] T. Tieleman and G. Hinton, "Lecture 6.5-rmsprop: Divide the gradient by running average of its recent magnitude," COURSERA: Neural Networks for Machine Learning. Accessed on: 20 January 2019 [Online]. Available: https://en.coursera.org/learn/neural-networks-deep-learning

[42] N. A. Weiss, *Introductory Statistics*. London, U.K.: Pearson. 2017.